DEEP SPACE COMMUNICATION AND NAVIGATION STUDY

Volume 3—System Considerations

May 1, 1968

FINAL REPORT

Contract No. NAS 5—10293

Contracting Officer: R. M. Keefe, NASA

Technical Monitor: J. E. Miller, NASA

Project Manager: J. S. Cook, BTL

Prepared by Bell Telephone Laboratories, Incorporated
Whippany Road, Whippany, New Jersey
On behalf of Western Electric Company, Incorporated
83 Maiden Lane, New York, New York
For National Aeronautics and Space Administration
Goddard Space Flight Center
Greenbelt, Maryland

ABSTRACT

This study provides a comparison of alternative means for high data rate communication (about 10^6 b/s) from deep space probes, and indicates the extent to which orbiting spacecraft can aid deep space navigation. Emphasis is on the communication problem. A special effort has been made to delineate practical and theoretical constraints on communication from a distance of 1 to 10 AU at microwave, millimeter, and optical frequencies (1 to 100 GHz and 20 to 0.2 microns wavelength), and to indicate promising avenues for extending the art.

The interrelationship between fundamental theory, device characteristics, and system performance has received particular attention in this study. Specific missions have been synthesized, and problems of visibility, Doppler variation, handover, acquisition, tracking, and synchronization have been investigated in order to discover the limitations imposed by practical system considerations.

This study was initiated and directed by Ira Jacobs.

CONTENTS
VOLUME 3

Volume 3. SYSTEM CONSIDERATIONS

Chapter 5. System Comparisons

Chapter 6. Tracking and Navigation Studies

Appendices

ILLUSTRATIONS
VOLUME 3

TABLES
VOLUME 3

CHAPTER 5. SYSTEM COMPARISONS

1. CANONIC MISSION – MARS ORBITER

In this section, a series of topics are considered which dictate how certain choices of orbital parameters affect orbital properties important to communication. The results are summarized here, with the details given in Appendix 10.

1.1 Fraction of Time a Mars Orbiter Is Occulted by the Planet

Under suitable simplifying assumptions, such as a cylindrical shadow, one can produce graphs of occultation fraction (defined as the fraction of an orbiter period for which Mars occults the orbiter) which are valid for at least several orbiter periods. These graphs demonstrate qualitatively the effects of varying the orbiter parameters and thus may be useful in preliminary mission design.

1.2 Fraction of Time Mars Is Within the Beam of the Earth Transmitter

This fraction is equal to the occultation fraction plus a small correction which is proportional to the distance subtended by the antenna beam at Mars and inversely proportional to the orbital radius of the orbiter.

1.3 Magnitude and Variation of Doppler Shift

The Doppler frequency shift is approximately periodic with a period of 2 years. The maximum fractional shift of 2.0×10^{-4} (which must be multiplied by the frequency to obtain the actual shift) occurs about 1-1/2 months after Earth-Mars opposition.

1.4 Visibility Conditions from a Tracker Satellite Situated at a Triangular Libration Point

An upper bound on the occultation fraction of 0.006 is obtained by neglecting the inclinations of the Mars orbit plane and the moon orbit plane to the ecliptic. This fraction represents only a few hours per month, which is negligible unless occultation occurs at times when continuous communication is necessary. This can be avoided by proper scheduling or redundant trackers.

1.5 Visibility Periods of a Mars Orbiter Relative to a Space Probe Approaching from Earth

Even for relatively small probe-Mars distances (as small as 0.01 AU), simple graphical results as in Section 1.1 can be used directly (and for even smaller distances they can be used with slight modifications) for preliminary mission design. A simple occultation criterion based upon the assumption of a conical shadow (and thus valid for any probe-Mars distance) is also given.

1.6 Payload Considerations for a Mars Mission

Several of the classes of possible missions are described, and references to some of the extensive recent literature are provided, in Appendix 10.

1.7 Visibility Conditions Between a Mars Landing Vehicle and a Mars Orbiter

This situation is similar to the one discussed in Section 1.5 except that an even simpler criterion for occultation can be given once the landing vehicle is on the Martian surface.

1.8 Visibility of a Mars Synchronous Satellite from an Earth Synchronous Satellite

It is demonstrated that there are periods of two or three months, two or three times per year when continuous communication is possible.

2. COMPARISON OF MICROWAVE SYSTEMS

2.1 Space Vehicle ERP

Over the past decade more than five orders of magnitude improvement in deep-space communication capability* has been achieved. This is illustrated in Table 50, where system parameters are given for Pioneer IV (1959), Mariner II (1962), Mariner IV (1965), and Voyager (1973).[1] The last column of this table gives the performance relative to Mariner IV. The 51.4-dB improvement in capability of Mariner IV relative to Pioneer IV was achieved by a 15.7-dB increase in spacecraft transmitter power, a 21.5-dB increase in spacecraft antenna gain (facilitated by an increase in system frequency from 960 MHz to 2290 MHz), and a 14.2-dB reduction in receiver noise occasioned by the introduction of a maser amplifier. For Voyager, a further 26.4-dB improvement is planned, consisting of 7-dB increased transmitter power, 8-dB increased antenna gain, 8-dB increased receiver aperture, and 3.4-dB reduced system noise temperature.†

It is pertinent to inquire into what further improvements might realistically be expected in the 2290-MHz DSIF system in the 1980 time period. Further significant improvements in receiver aperture (see Section 2.2) or noise temperature are not anticipated. The improvements then will come largely from increased space vehicle effective radiated power.

In Figure 115, contours of constant ERP are shown on a plot of transmitter power (in dBW) versus antenna gain (in dB). The transmitter power and antenna gain of the four systems compared in Table 50 are shown by dots on this figure. The dashed straight line drawn through the three uppermost points indicates that, although power and antenna gain are both increasing to achieve increased ERP, gain is increasing somewhat more rapidly than power; viz., $G \sim P^{5/7}$.

Although there are generally constraints which limit the choice of power (availability of space-qualified tubes and heat-dissipation considerations) and antenna size (shroud dimensions and pointing considerations), it is still of interest to determine the relation between power and gain to achieve a given ERP with minimum weight. The weight required to achieve a given radiated power in the vicinity of Mars (assuming solar cell prime power) is shown in Figure 116, and the corresponding weight for a deep-space mission using nuclear primary power is shown in Figure 117. Although tube weight increases as $P^{1/2}$, the principal contribution to the weight is made by the primary power source, and this weight increases linearly with P. Consequently, the weight W_P required to achieve a given radiated power P is given by

$$W_P = W_1 + w_p P \qquad (1)$$

where W_1 is a fixed weight, and w_p (in pounds per watt) is the incremental weight associated with an increase in power. Assuming solar cell prime power at 1.7 AU from the sun and an overall efficiency of 40 percent, $w_p \approx 0.6$ pounds per watt of radiated power.

The relation between weight and antenna gain (or antenna diameter) cannot be fit by a simple functional relation. This is illustrated in Figure 118 where the results of Chapter 1, Section 3 are used to plot spacecraft antenna weight gain for space-craft antennas at 2.3 GHz and 8 GHz. The weight is shown as a function of diameter in Figure 119; the weight is essentially independent of frequency for diameters below 5 meters, but as the diameter is increased the weight increases more rapidly for the higher frequency.

If attention is restricted to a limited region of the curves, the curves in Figures 118 and 119 may be approximated (albeit not too well) by straight lines. This is illustrated by the dashed line in Figure 118 which indicate that, at 2.3 GHz, the weight of the spacecraft antenna increases as $G^{0.6}$ (or $D^{1.2}$). The fact that the weight increases more slowly than the antenna area is indicative of the fact that most of the weight resides in the supporting structure rather than in the dish itself.

If antenna weight is assumed to be given by

$$W_a = W_2 G^{0.6} \qquad (2)$$

it follows that the combined weight of antenna and power is given by

$$W = W_a + W_P = W_2 G^{0.6} + W_1 + w_p P \qquad (3)$$

For a fixed ERP = PG, the optimum choice of P and G to minimize W is given by

$$P = \left(\frac{0.6 W_2}{w_p}\right)^{\frac{5}{8}} (\mathrm{ERP})^{3/8} \qquad (4)$$

$$G = \left(\frac{w_p}{0.6 W_2}\right)^{\frac{5}{8}} (\mathrm{ERP})^{5/8} \qquad (5)$$

and the optimum weight is given by

$$W_{min} = W_1 + 1.9\, W_2^{5/8}\, w_p^{3/8}\, (\mathrm{ERP})^{3/8} \qquad (6)$$

The above results indicate that as ERP increases the weight of the spacecraft communications increases as

---

*Capability is measured here in terms of the signal-to-noise ratio at a given range and bandwidth. Information rate increases linearly with SNR and range increases as $(SNR)^{1/2}$.

†The additional sky noise, when Mars is within the receiving beam, is given by $T\Omega_m/\Omega_b$ where T is the equivalent black body noise temperature ($T \approx 200$ °K) of Mars, Ω_m is the solid angle subtended by Mars ($\Omega_m \approx 10^{-6}$ sr), and Ω is the solid angle beamwidth of the receiving antenna ($\Omega \approx 3(10)^{-6}$ sr), where we have assumed $\Omega_m < \Omega$. (If the converse is true, the noise temperature is T.) Thus, the additional noise is of the order of 1°K and is negligible.

Table 50

DEEP SPACE COMMUNICATION SYSTEMS

		f (GHz)	P_T (watts)	G_T (dB)	D_R (feet)	T (°K)	Δ* (dB)
Pioneer IV	(1959)	0.96	0.27	2.5	85	1450	-51.4
Mariner II	(1962)	0.96	3	19	85	250	-16.8
Mariner IV	(1965)	2.29	10	24	85	55	0
Voyager	(1973)	2.29	50	32	210	25	+26.4

$$*\Delta = 10 \log \frac{(P_R/T)}{(P_R/T)_{\text{Mariner IV}}}$$

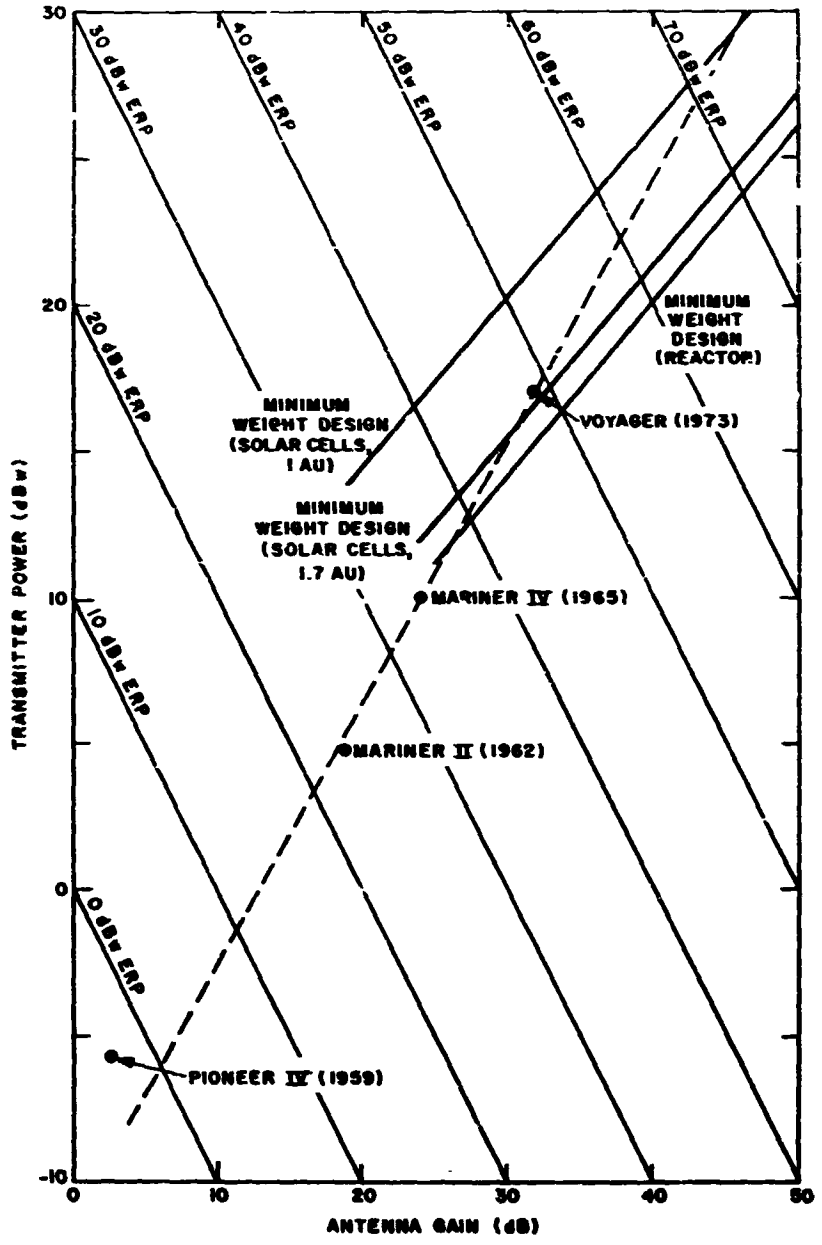


Figure 115. ERP of deep space communications systems

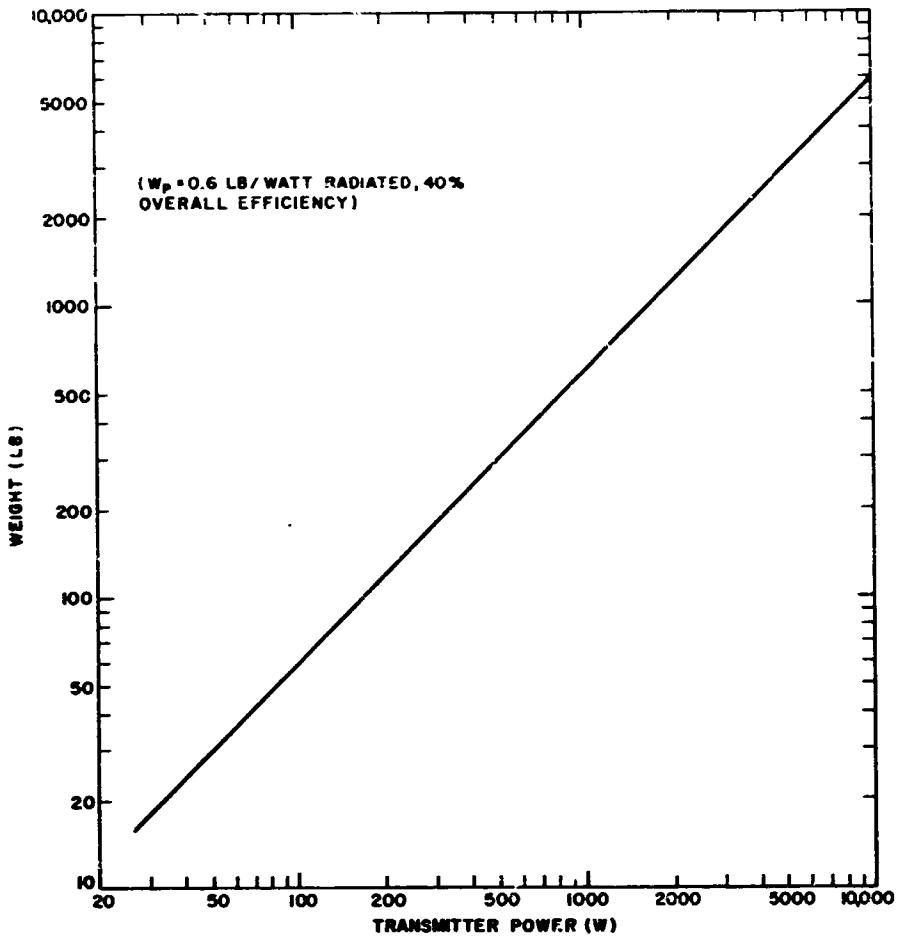


Figure 116. Weight of transmitter vs. power output for solar cells in the vicinity of Mars

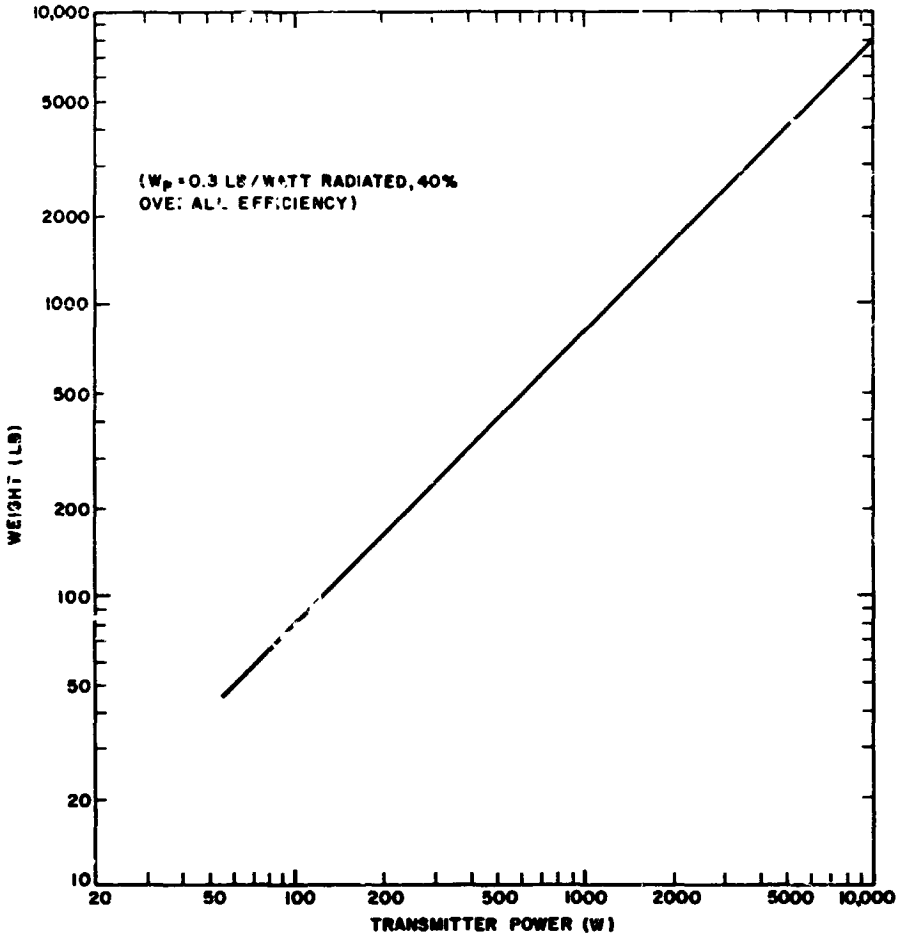


Figure 117. Weight of transmitter vs. power output for reactor power source

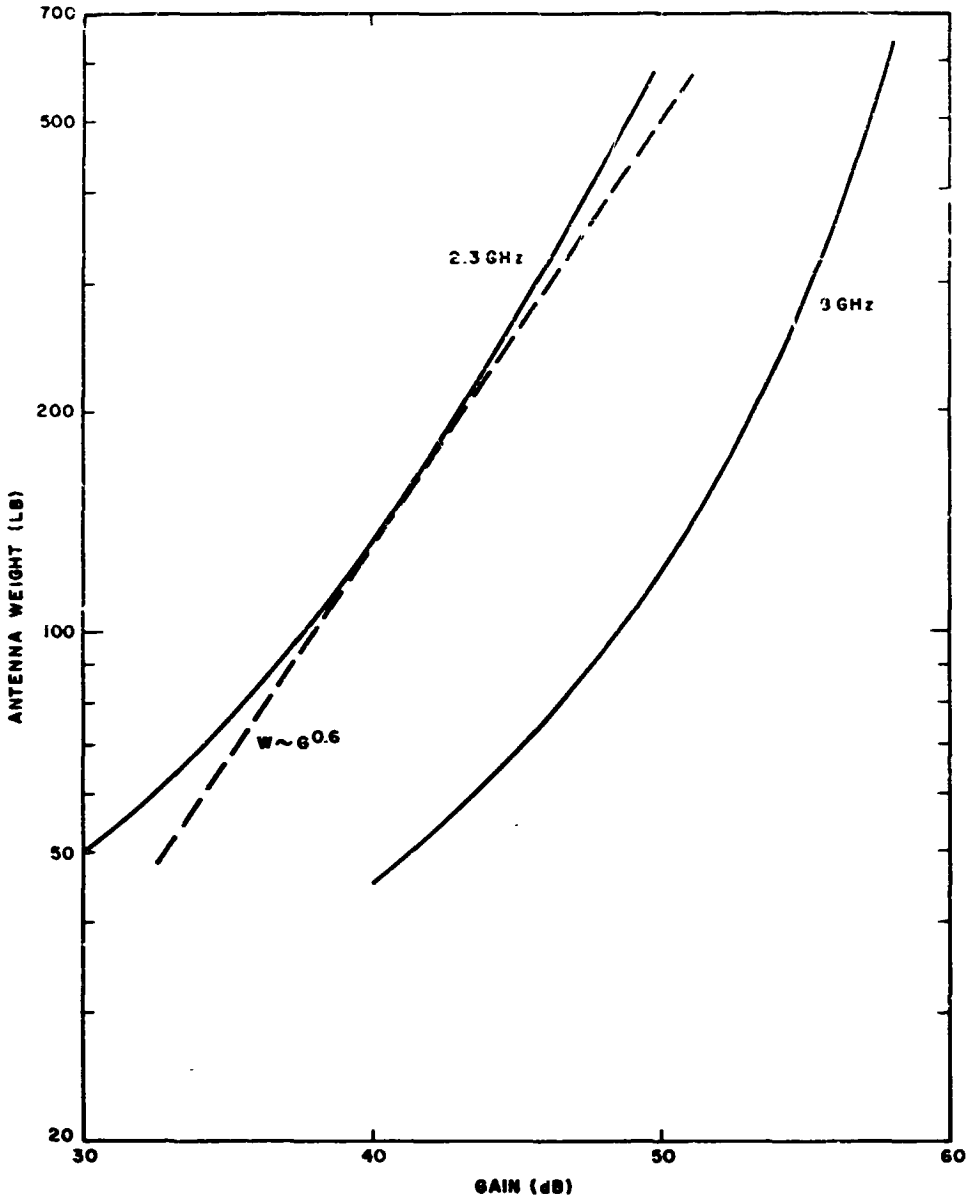


Figure 118. Space vehicle antenna weight required to achieve a given gain

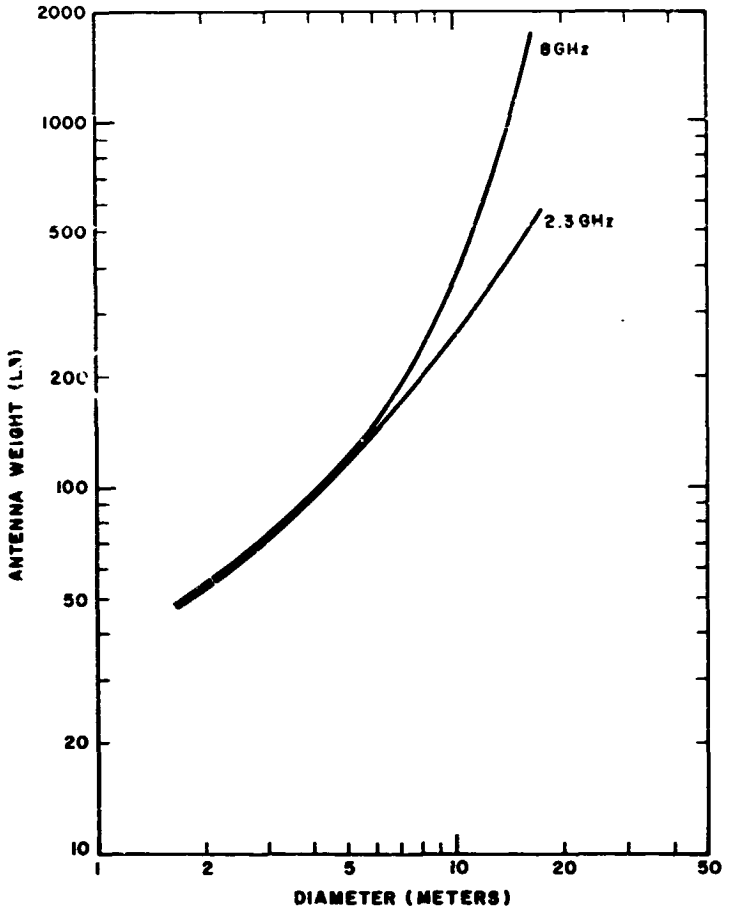


Figure 119. Space vehicle antenna weight required to provide a given diameter

$(ERP)^{3.8}$. Thus, to go from a 50-dBW ERP (Voyager) to 70 dBW would require an increase in weight by a factor of about 6.

The above results also indicate that, as ERP increases, the "optimum" design has gain increasing more rapidly than power; specifically, $G \sim P^{5/3}$. The empirical data in Figure 115 indicated that $G \sim P^{8/7}$. However, as experience is gained with large lightweight space erectable antennas, the higher exponent found above may be more appropriate.

The optimum power and antenna gain, as obtained from Equations (4) and (5), is shown by the dark lines in Figure 115 for the cases of

1. Solar cells in the vicinity of the Earth (1 AU)
2. Solar cells in the vicinity of Mars (1.7 AU)
3. Reactor prime power.

As the weight of prime power (w_p) increases, the optimum design to achieve a given ERP uses relatively less power ($P \sim w_p^{-5/8}$) and more gain ($G \sim w_p^{5/8}$). The weight factor, w_p, for a reactor is about 4 times that of solar cells in the vicinity of Earth. Thus reactors will require less weight than solar cells for missions extending beyond 2 AU from the Sun.

It is interesting to note from Figure 115 that the Voyager parameters correspond to a minimum-weight design although, because of the approximate nature of the models, the correspondence is probably only fortuitous.

In Figure 120, the weight required to achieve a given ERP is shown for the three cases noted above, assuming the minimum weight design. The weight includes antenna, prime power, power supply, and transmitting tube.

2.2 Communication Performance at 2.3 GHz

The information rate achievable in a deep-space communication system is given by

$$H = \frac{PGA}{4\pi R^2 \, kT(E/N_0)} \qquad (7)$$

where P = the transmitter power

G = the gain of the transmitting antenna

A = the effective area of the receiving antenna

R = the range

k = Boltzmann's constant ($k = 1.38(10)^{-23}$ joule/°K)

T = the system noise temperature

E/N_0 = the ratio of energy-per-bit to noise spectral density required to achieve a desired error probability (see Chapter 1, Section 6).

It was noted in Chapter 1, Section 6 that the best coherent binary communication system (phase shift keying) requires $E/N_0 \approx 10$dB to achieve an error probability of 10^{-5}, but by the use of larger alphabet modulation and/or coding systems, a 5-dB reduction can be achieved. For the purpose of the performance calculations in this chapter, $E/N_0 = 10$ dB will still be used to allow both for losses in the transmitter and receiver and for some margin. The intent is to provide a basis of comparison for systems operating in the various frequency bands, rather than to make a precise evaluation of the information rate at each frequency.

In Figure 121, information rate is plotted as a function of range for ERP = 50, 60, 70, 80 dBW, assuming DSIF receiving parameters [$A = 2.25(10)^3 m^2$, T = 25°K]. To achieve an information rate of 10^6 bits per second from 1 AU requires an ERP of 57 dBW. This could be achieved, for example, with the same 50-watt (17 dBW) tube as in Voyager and a 5.8-meter (19-foot) antenna which appears well within the state of art. To achieve 10^6 bits per second from 10 AU would require an ERP of 77 dBW which could be achieved (see Figure 120) with a 500-watt transmitter (27 dBW), and an 18.2-meter (60-foot) antenna. In Figure 118 this antenna is estimated to weigh 600 pounds. In Figure 121 the weight of power plus antenna is estimated at 1000 pounds. Thus, it would appear that present microwave technology is sufficient to achieve 1 Mb/s from a distance of 1 AU, and that the achievement of 1 Mb/s from a distance of 10 AU is consistent with reasonable estimates of future space transmitters and antennas.

2.3 Effect of Increasing Frequency

Frequency does not appear explicitly in Equation (7), but it appears implicitly in the factors P, G, A, and T. As noted in Chapter 1, Section 2, although for a particular tube design P scales as $1/f^2$, in practice radiated power will be limited by prime power and heat dissipation considerations. Thus, it is reasonable to assume that transmitter power is independent of frequency over the microwave region.

Under cloud cover and light rain conditions, the sky temperature at 8 GHz is about 10°K for a 30-degree elevation angle (see Figure 39). If the receiver is sited in a location where heavy rain is improbable, it may be assumed that there is essentially no noise penalty in increasing the frequency from S to X band.

Thus, the effect of increasing the frequency within the microwave band is determined by the frequency dependence of G and A. In Figure 122 the gain of a spacecraft antenna of fixed weight is shown as a function of frequency (see Chapter 1, Section 3). For a 200-lb space vehicle antenna, gain increases as $f^{1.6}$. Note that the relative advantage of increasing frequency is greater when one is constrained to a low weight than when one is allowed a large weight. This is readily understood since low weight constrains one to small antennas.

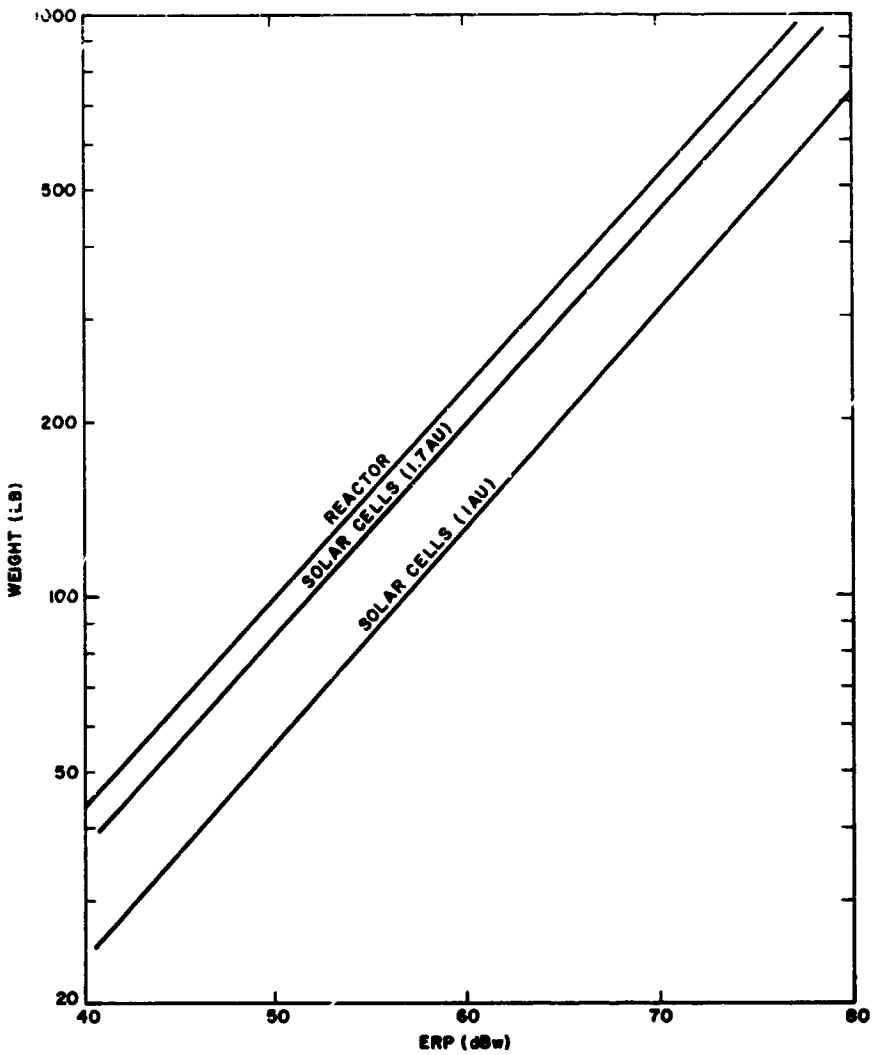


Figure 120. Weight required to achieve a given ERP

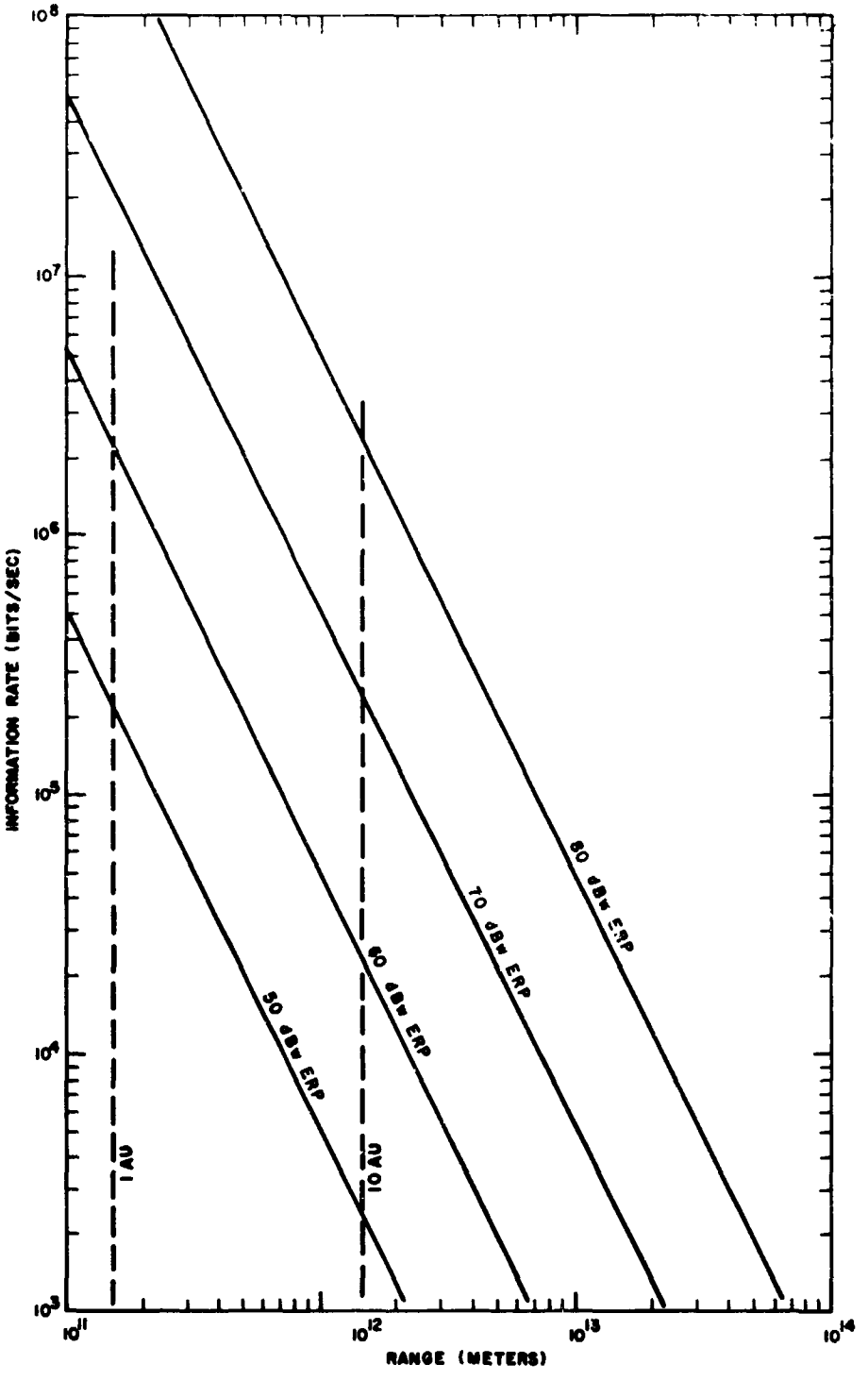


Figure 121. Information rate achievable with DSIF

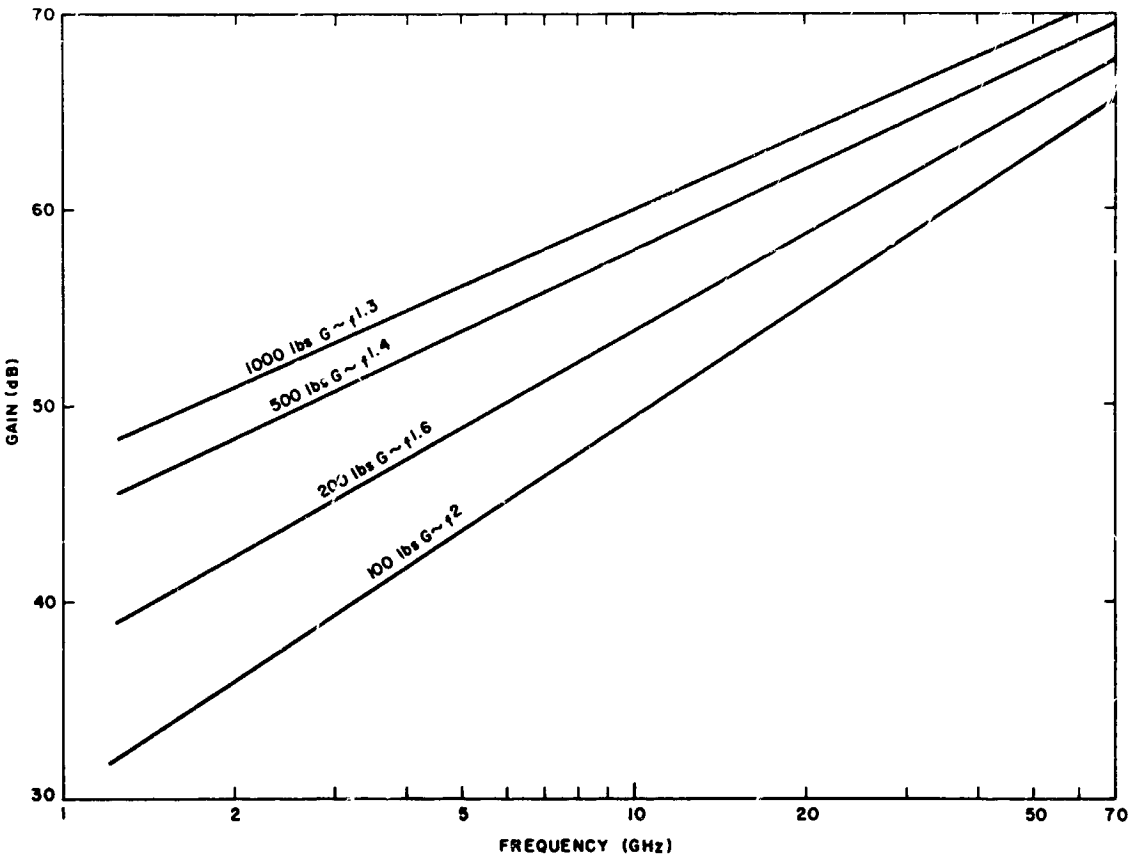


Figure 122. Frequency dependence of space vehicle antenna gain for fixed weight

In the case of the receiving antenna, cost rather than weight is the appropriate parameter to be fixed as frequency is varied. The antenna cost model, developed in Chapter 1 Section 5, was used to compute gain versus diameter curves for fixed cost. These are shown in Figures 123 through 126 for costs of $1, $5, $10, and $20 million. For a fixed cost and frequency there is a diameter which achieves maximum gain. Above this diameter a gain-limited antenna cannot be achieved with the given cost. Below this diameter, although the antenna is gain-limited, gain falls off as diameter decreases.

In Figure 127 the maximum gain (obtained from Figures 123 to 126) is shown as a function of frequency for fixed cost. The results are fit quite well by two sets of straight lines:

1. For 8 GHz and below, $G \sim f^{1.6}$. Consequently, $A \sim f^{-0.4}$.
2. For 16 GHz and above, $G \sim f^{1.1}$. Consequently, $A \sim f^{-0.9}$.

The fact that effective area decreases more rapidly at the higher frequencies is a consequence of manufacturing tolerances limiting antenna sizes at these frequencies.

Thus, in going from 2 to 8 GHz, with a 200-lb spacecraft antenna, the gain may be increased from 42 to 52 dB (Figure 122). For a $10 million ground antenna, the receiving gain may be increased from 60 to 70 dB (Figure 127) which corresponds to a 2-dB loss in effective area. Thus the GA product is increased by 8 dB, and, therefore, an X-band system is expected to have a factor 6 more information rate capability that an S-band system with corresponding spacecraft weight and ground terminal cost.

Continuing with the above example, if the frequency is increased to 16 GHz, the spacecraft antenna gain may be increased an additional 5 dB to 57 dB, but the ground antenna gain (Figure 127) may be increased only to 74 dB which corresponds to a 2-dB loss in effective area relative to the X-band case. Thus, the GA product is increased an additional 3 dB. However, at 16 GHz, the additional sky noise (under cloud and light rain conditions $T_{sky} \approx 25°K$) should more than negate the advantage of improved GA product.

3. COMPARISON OF MILLIMETER SYSTEMS

The spacecraft antenna weight and ground antenna cost data (Figures 122 and 127) presented in the previous section cover the frequency range from 1 to 100 GHz, and hence include the two "atmospheric windows" at 35 and 94 GHz. The above results have been combined to give the product of transmitting antenna gain and receiving antenna effective area as a function of frequency for fixed transmitting antenna weight and ground receiving antenna cost. The results are shown in Figure 128 where the curves are labelled by the spacecraft antenna weight and the ground terminal cost.

If transmitter power and receiver power were independent of frequency and there were no atmospheric attenuation, then the ordinate of the above curve would be proportional to communication rate. These assumptions are valid in the microwave (1 to 8 GHz) region barring heavy rainfall (see Chapter 1, Section 4), and hence there is an advantage in increasing frequency to X-band as noted in the previous section. It is of interest to note from Figure 128 that the relative advantage (i.e., the slope of the curves) is greater the smaller the spacecraft weight.

In the millimeter region, the slope of the curves diminishes appreciably because manufacturing tolerances limit consideration to smaller antennas. The effect is less pronounced at the lower weights, primarily because the lower weight already provides an antenna size constraint.

The curves in Figure 128 are monotonic increasing, so that millimeter wavelengths would offer an advantage if there were no frequency-dependence of transmitter power, receiver noise, and attenuation. Unfortunately, although the first is approximately true, neither the second nor third assumptions are applicable in the millimeter band.

As noted in Chapter 2, Section 1, coupled cavity TWTs have been built with output powers in the range of several hundred watts to several kilowatts throughout the millimeter band, and with efficiencies comparable to those achieved at microwaves. Although no such tubes have as yet been space qualified, there is little reason to expect that this could not be done. The biggest problem, particularly when relatively high powers are desired, is heat dissipation. Because of the smaller dimensions this is a more serious problem at millimeter wavelengths and consequently some additional weight may be required for millimeter tubes relative to microwave tubes. For the purpose of performance calculations here, however, it will be assumed that power is limited by the weight of the prime power, and that there is no frequency dependence of radiated power throughout the microwave and millimeter region. To achieve this result would require considerable development effort at the millimeter wavelengths.

Although it may be assured (somewhat optimistically) that millimeter systems will not suffer a power penalty relative to microwave systems, there is a penalty associated with atmospheric attenuation (see Chapter 1, Section 4 and Chapter 2, Section 2). In Table 51 the atmospheric propagation losses when observing at 30° elevation angle, are given at frequencies of 2, 8, 16, 35 and 94 GHz due to various atmospheric conditions. It is clear that millimeter systems cannot operate through heavy rain, but by appropriate siting and diversity, heavy rain may be avoided. It will be assumed, however, that the system must be capable of operating in the presence of light rain. In Table 52 the attenuation under these conditions is given together with the additional sky noise (T_{atm}) contributed by this attenuation. Total system noise temperature is assumed to be given by $T_{atm} + 25°K$ which assumes that low-noise

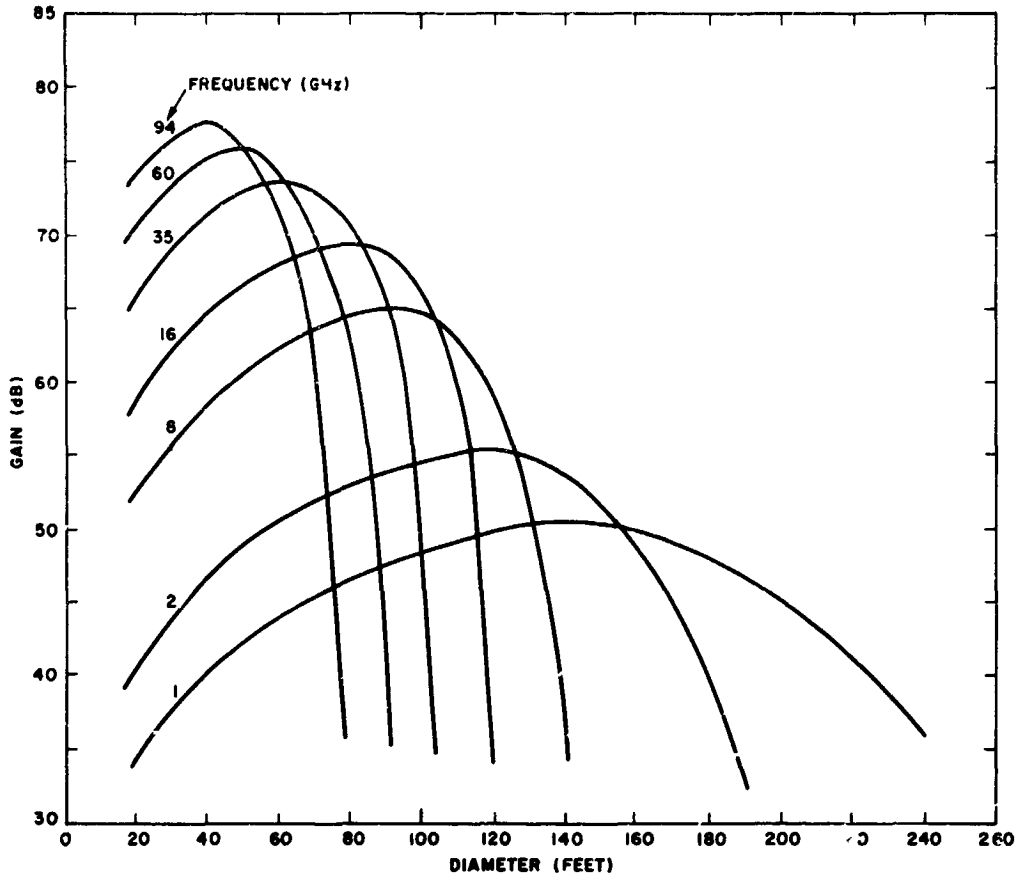


Figure 123. Antenna gain vs. diameter for several frequencies (fixed cost: $1,000,000)

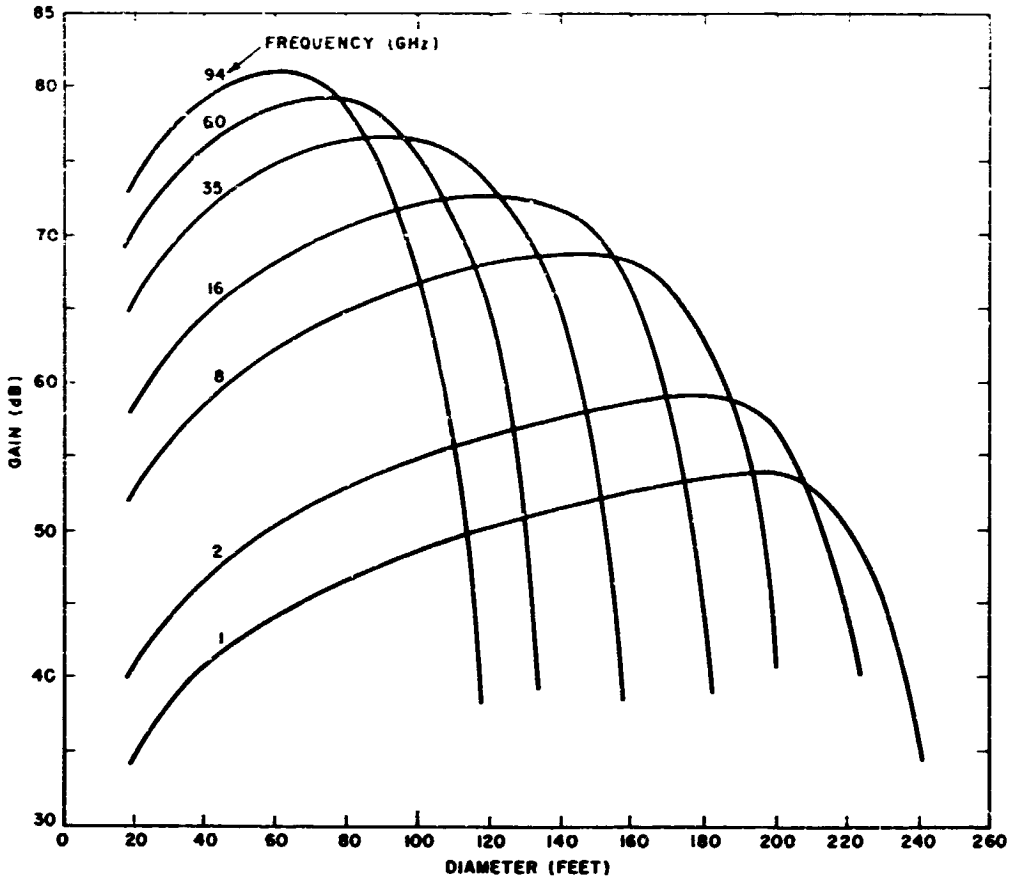


Figure 124. Gain vs. diameter for several frequencies (fixed cost: $5,000,000)

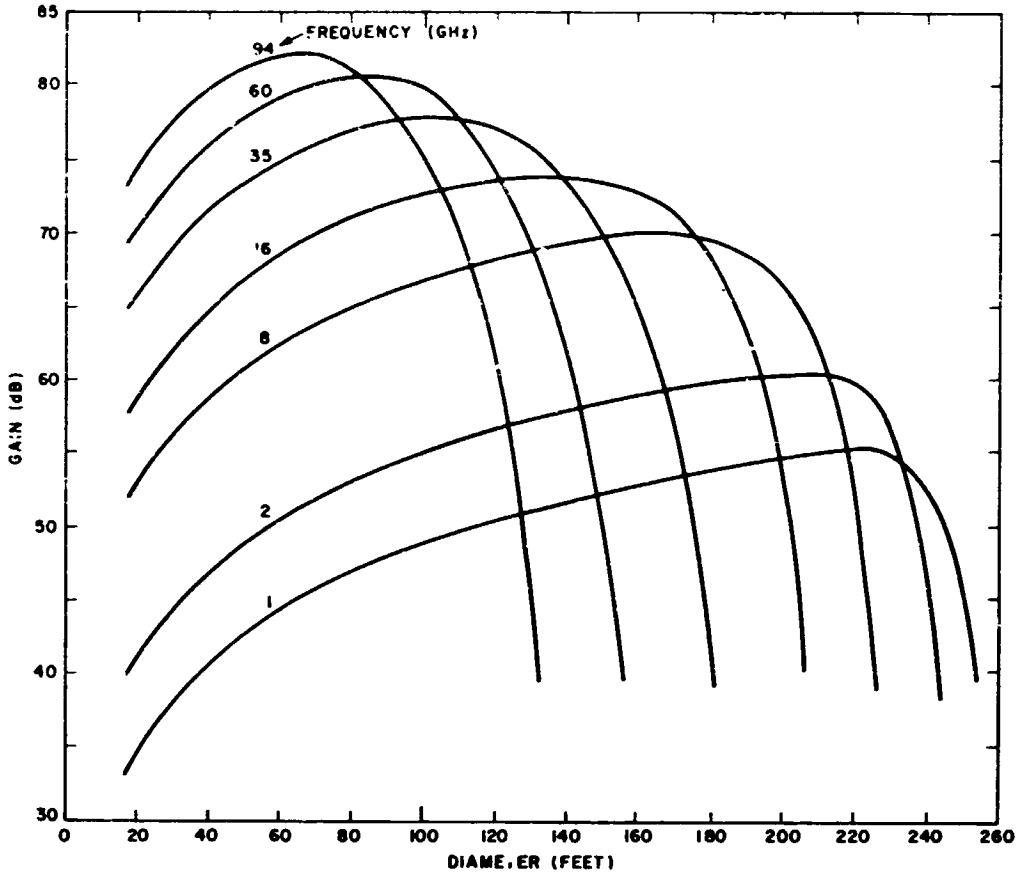


Figure 125. Gain vs. diameter for several frequencies (fixed cost: $10,000,000)

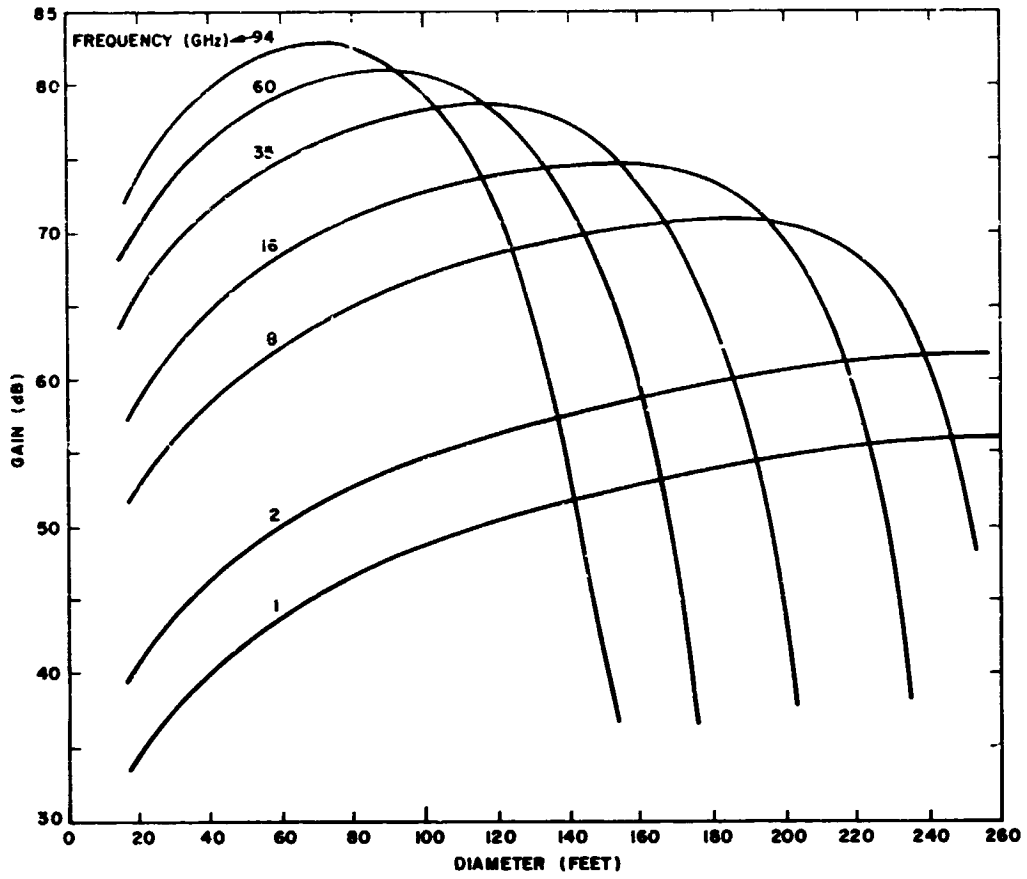


Figure 126. Gain vs. diameter for several frequencies (fixed cost: $20,000,000)

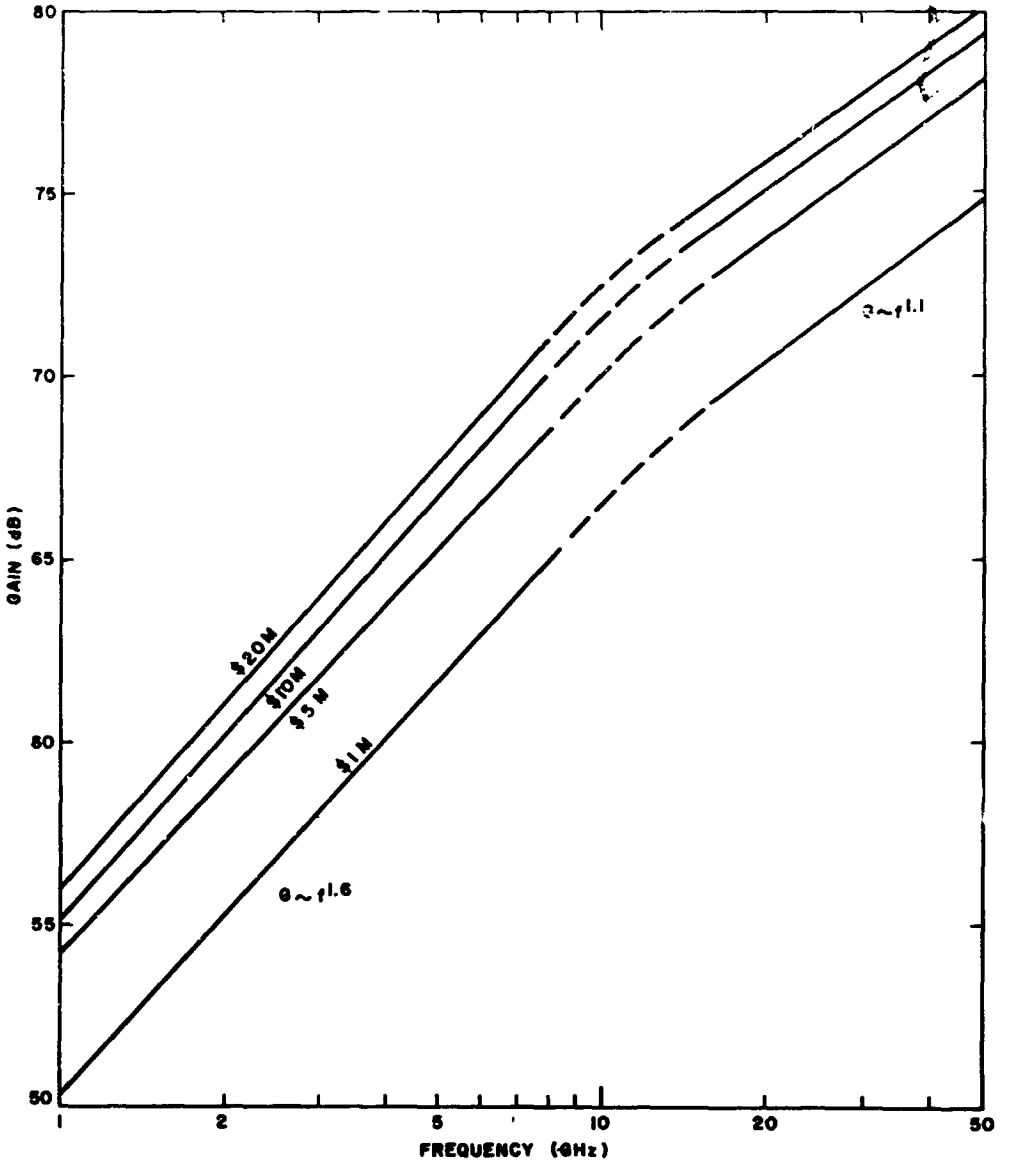


Figure 127. Maximum gain of large fixed-cost ground antennas

Table 51

ATMOSPHERIC PROPAGATION LOSSES (IN dBs)

	2 GHz	8 GHz	16 GHz	35 GHz	94 GHz
Clear atmosphere	0	0	0.2	0.5	2.4
Dense clouds	0	0.1	0.3	2	9
0.1 in/hr rain	0	0.1	0.6	3	11
1 in/hr rain	0	1.8	12	30	60

Table 52

PERFORMANCE DEGRADATIONS CAUSED BY ATMOSPHERIC LOSS UNDER CONDITIONS OF .1 in/hr RAINFALL OR VERY DENSE CLOUD COVER

	2 GHz	8 GHz	16 GHz	35 GHz	94 GHz
Att (dB)	0	0.1	0.8	3.5	13.4
T_{atm} (°K)	0	7.2	24	130	285
T	25	32.2	49	155	310
10 log T/25	0	1.1	2.9	7.9	11
Degradation (dB)	0	1.2	3.7	11.4	24.4

maser receivers are available throughout this band (see Chapter 2, Section 5, which indicates that the assumption is reasonable up to 35 GHz but questionable at 94 GHz).

It is seen from Table 52 that there is an appreciable degradation at the millimeter wavelengths associated with the increased sky noise.

In Table 53 the results of Figure 128 and Table 52 are combined to give the signal-to-noise ratio (and hence the information rate) on a relative decibel scale for the frequencies considered above and for the various combinations of space antenna weight and ground antenna cost contained in Figure 128.

To convert the relative performance data in Table 53 to information rate H, it is necessary to assume that a transmitter power P, a range R, and a performance measure of the modulation system E/N_o. For P = 100 watts, R = 1 AU, and $E/N_o = 10$ [H is proportional to $P/(R^2\ E/N_o)$], the 0 dB entry in the table correponds to $2.5(10)^5$ bits per second. Thus, for example, the use of a 100 pound space antenna and a $1 million ground antenna would achieve an information rate of $2(10)^6$ bits per second at f = 8 GHz (assuming again P = 100 watts, R = 1 AU, $E/N_o = 10$).

Several interesting conclusions may be drawn from Table 53.

1. There is an appreciable advantage (6 to 9 dB) in going from 2 to 8 GHz.

2. The performance at 16 GHz is essentially the same as at 8 GHz under the assumed light rain conditions. Note, however, from Table 51 that there is considerably more degradation at 16 GHz than at 8 GHz under heavy rain conditions.

3. The performance relative to that at 8 GHz degrades appreciably at the millimeter frequencies. Note, however, that for the lighter spacecraft antennas, the performance at 35 GHz is better than that at 2 GHz. This conclusion must be tempered, however, by the extreme sensitivity to weather conditions and the current unavailability of space qualified millimeter tubes. It is generally true, however, that the millimeter frequencies are relatively more attractive when there is a tight constraint on spacecraft weight.

Table 53

RELATIVE SYSTEM PERFORMANCE (dB)

Space Antenna Weight (lbs)	100	100	200	500
Ground Antenna Cost (dollars)	10^6	10^7	10^7	10^7
f = 2 GHz	0	4	11	17
f = 8 GHz	9	13	18	23
f = 16 GHz	10	15	18	22
f = 35 GHz	6	11	13	16
f = 94 GHz	0	4	5	7

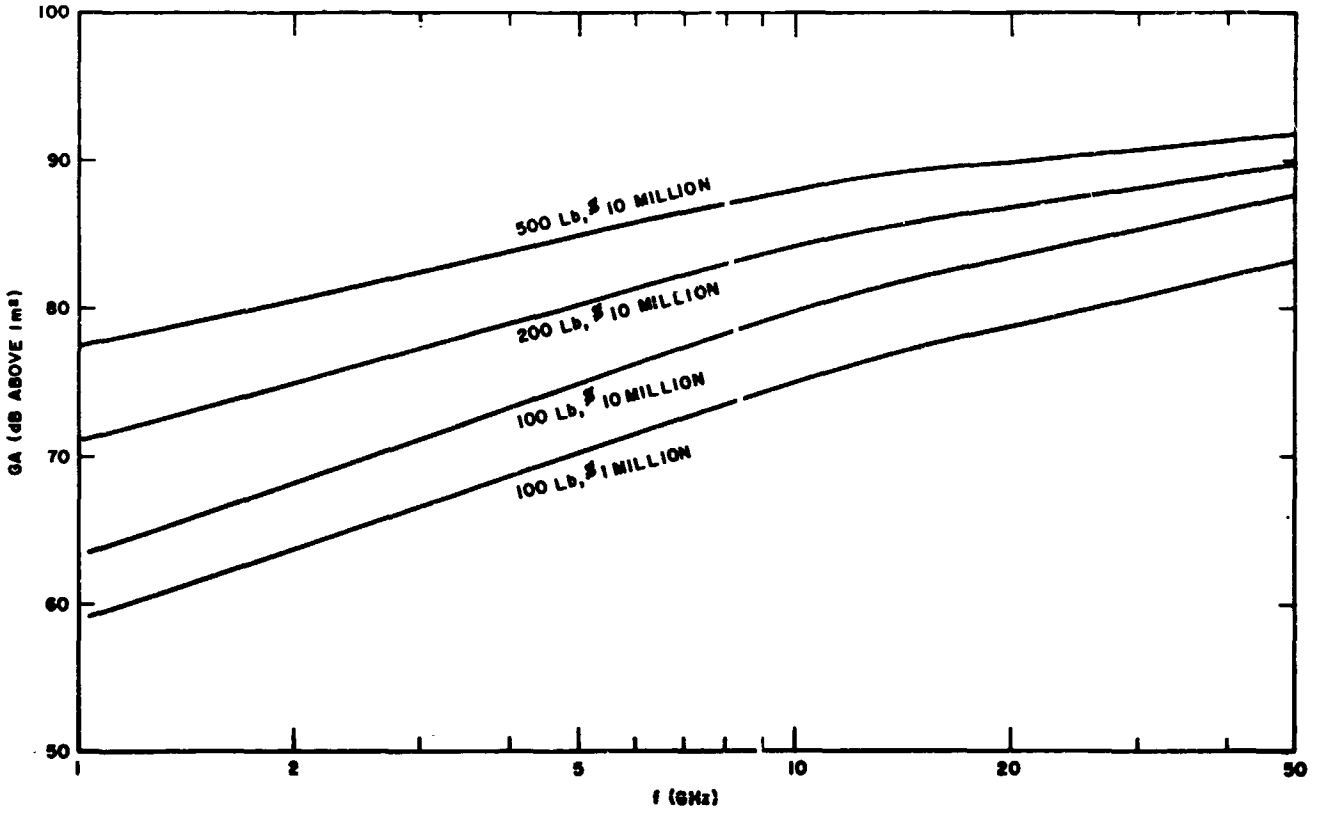


Figure 128. Frequency dependence of product of gain of fixed-weight transmitting antenna and area of fixed-cost receiving antenna

4. MILLIMETER SYSTEMS WITH SATELLITE RECEIVER

As noted in the previous section, the margins required for atmospheric effects may make millimeter wavelengths less attractive than S band for communication from a deep space vehicle to an Earth receiver. It is of interest then to consider a deep-space millimeter communication system in which the receiving terminal is located on an Earth satellite rather than on ground, with communication from the satellite to the ground via S band. This, of course, permits the use of frequencies outside the atmospheric windows for transmission from the space vehicle to the Earth satellite.

In Figure 129 the gain and diameter of a 200-pound spacecraft antenna (as obtained from the results of Chapter 1, Section 3) are shown as a function of frequency. At 2.3 GHz, the gain is 44 dB corresponding to an antenna diameter of 9.1 meters (30 feet). At 100 GHz, the gain is 70 dB corresponding to an antenna diameter of 4.3 meters (14 feet). For a fixed weight spacecraft antenna, the gain is proportional to $f^{1.6}$ which corresponds to the diameter being proportional to $f^{-0.8}$.

It is not possible to assign a simple functional relation for the frequency dependence of transmitter power and receiver noise temperature. For a given tube design, transmitter power tends to scale as $1/f^2$. On the other hand, efficient (40 percent) extended interaction TWTs are now being designed with output powers of the order of 100 watts (see Chapter 2, section 1). Although present powers are at least an order of magnitude below those achievable at S band, in the 1980 time frame the radiated power may be determined largely by prime power considerations and may be essentially independent of frequency in the microwave and millimeter region.

Similarly, although there are presently few masers operating at millimeter wavelengths (see Chapter 2, Section 6), there is no inherent reason why a millimeter system operating outside the atmosphere could not ultimately achieve the same low-noise performance as is presently achieved at S band. Unfortunately, however, because of atmospheric absorption, there has not been a strong reason to develop extremely low noise millimeter receivers.

In Figure 130, the satellite receiver antenna diameter required to achieve the same communication rate as a 64-meter (210-foot) S-band receiver is shown as a function of frequency for two cases.

1. Transmitter power divided by receiver noise temperature is the same at millimeter wavelengths as at S band.
2. This ratio is 10 dB poorer at the millimeter wavelengths than at S band.

In both cases the transmitting antenna gain is assumed to be given by Figure 129.

Under the optimistic assumption of Case 1, a 3.2-meter (10-foot) receiving antenna would be required at 100 GHz. Under the more realistic assumption of Case 2, the receiving antenna diameter is 10.1 meters. However, a 10-meter Earth satellite antenna, good at 100 GHz, is outside the range of presently contemplated design.

The above results indicate that a millimeter system with a satellite receiver would require extensive development efforts in the areas of space transmitters, low-noise space receivers, and high-gain (70 to 80 dB) space antennas just to equal the performance of an S-band system with the present DSIF receiver. The prospects are rather remote for obtaining performance at millimeter wavelengths appreciably better than S band.

5. COMPARISON OF GROUND VERSUS SATELLITE RECEIVER FOR OPTICAL SYSTEMS

There is a serious question as to whether the receiving site for an optical communication system should be located on earth or on a satellite outside the Earth's atmosphere. In the latter case, communication from the satellite to the Earth would be at S band where atmospheric effects may be neglected. The purpose of this section is to review and evaluate (as well as existing information allows) the arguments for and against the use of satellite receivers.

5.1 Siting of Ground Receivers

A general-purpose deep-space receiving network requires at least three ground terminals to assure continuous communication while the space vehicle is in view of the earth. Optical receiving sites should be located where atmospheric effects are minimized. A recent study[2] of siting for an Optical Communication Experimental Facility (OCEF) has pointed out that, although siting an optical ground terminal is similar to that of siting an astronomical observatory, there are important differences. (See also the discussion in Chapter 3, Sections 4.1 and 5.1.) Stated most simply, an observatory is concerned with maximizing both the duration and quality of the best seeing conditions, whereas an optical communication site is concerned with minimizing the extent and duration of poor conditions. The two are not synonymous, although both generally require location at high altitude in a dry region. The OCEF study restricted attention to sites above 6000 feet altitude to avoid principal atmospheric dust and pollution, and below 9000 feet because of physiological effects. For an operational site (as opposed to an experimental facility), it is feasible to consider a self-sustaining unattended site with communications to a control center by microwave relay, or perhaps by a microwave Earth satellite system. Although a site, once constructed, may in principle be largely unattended, the difficulties and cost of construction argue for sites with reasonable accessibility.

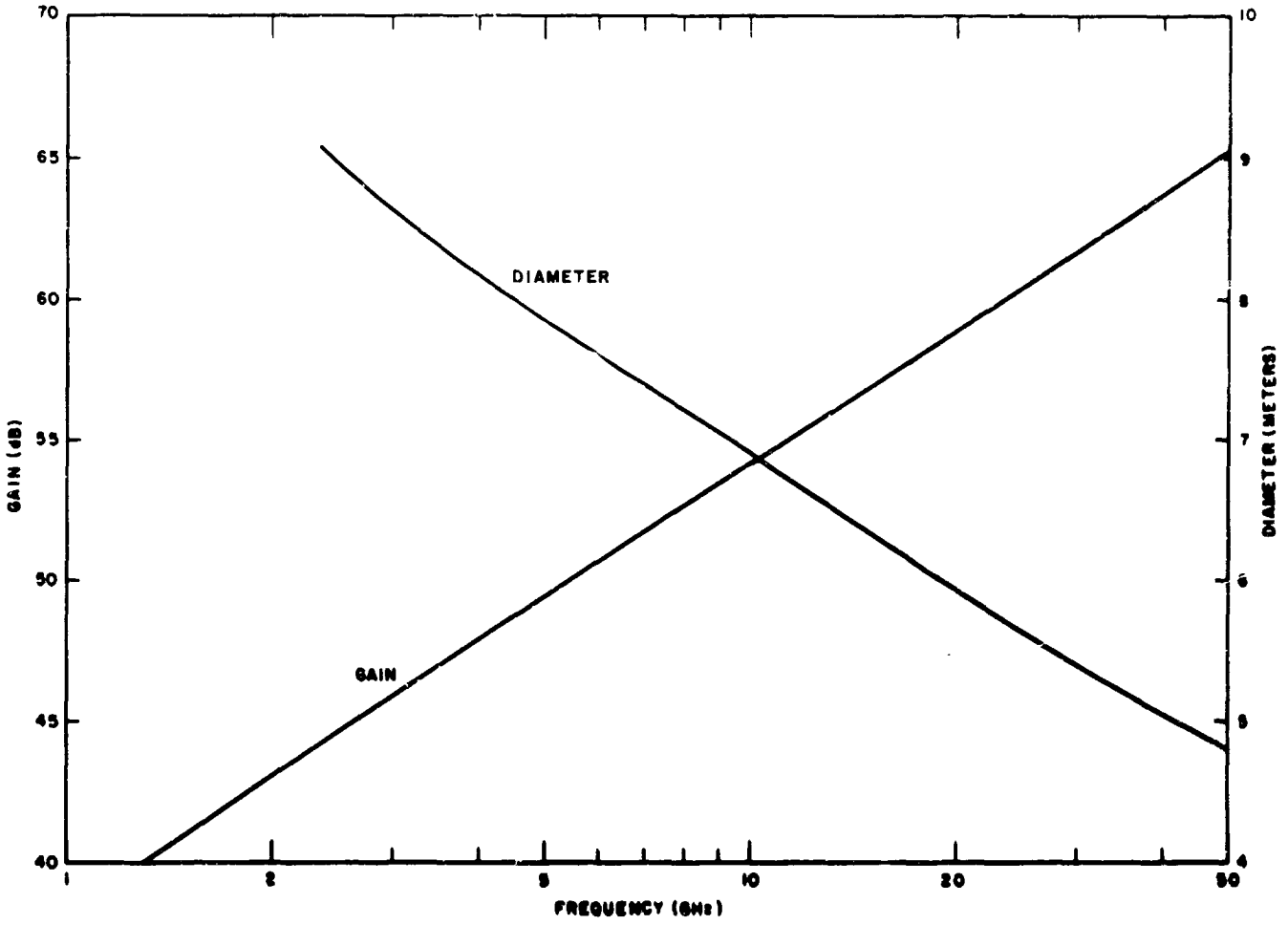


Figure 129. Gain and diameter of a 200-lb space vehicle antenna

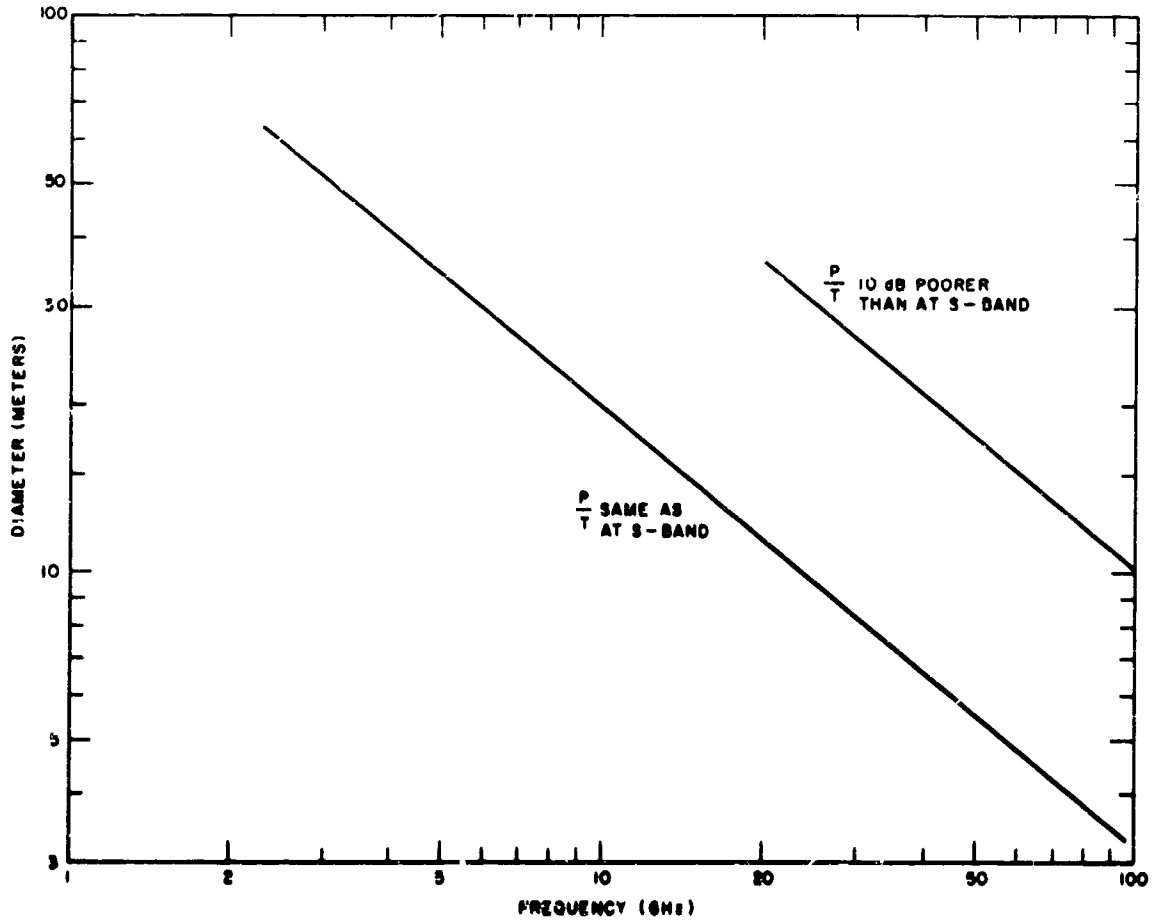


Figure 130. Satellite receiving antenna required to achieve same communication rate as S-band system with 64 m ground antenna

The OCEF study[2] recommended eight sites, all in the southwest (the study was restricted to the continental United States), but "no site with unusually high year-round reliability has been found." Outside the continental United States it would appear that both available data and available sites (considering political restrictions) are even more limited.

The single most important factor in determining the suitability of a particular site is cloud cover. Water vapor extinction coefficients are of the order of 500 $(db/km)/(gm/m^3)$ at 10 microns and 1200 $(db/km)/(gm/m^3)$ at 0.5 micron.* Thus a cloud with a water vapor content of 0.1 gm/m^3 and an extent of 1 km will give attenuations of the order of 50 db at 10 microns and 120 db at 0.5 micron. Certainly at optical frequencies it is completely impractical to provide sufficient power to overcome this attenuation. It is necessary therefore to find sites at which probability of cloud cover is small, and then, perhaps, use diversity to further improve reliability.

Data[2] exist on the average fractional portion of the sky which is obscured by clouds on a monthly basis, but the data do not appear adequate to determine quantities such as correlation of cloud cover at widely separated areas and the distribution of duration of outages and time between outages. Unlike heavy rain, which is very localized, "widespread high-altitude cloud is often distributed over large portions of the continent. In general, knowledge of such formations is difficult to accumulate so that little data is available."[2]

To illustrate the variability of the data, consider some results (quoted by Kalil[3]) on the percent of time lost due to clouds at Baker-Nunn sites. These data are reproduced in Table 54. In South Africa, from May through September of 1962, the monthly values did not exceed 8 percent. In 1963, May, June, and July were all above 15 percent. For the ten sites and the two years quoted, there was no example of a site which did not have at least one month in which the percentage of time was greater than 50 percent. Although there were several months which were good at particular sites (June, July, and August in Peru), there were months (notably January) at which none of the sites were reliable.

Although a study has by no means been made of all the available data, it appears that gross average cloud cover data can serve only to pinpoint particular sites at which more extensive measurements (e.g., monitoring solar radiation) should be performed. However, even if such data are taken, the variability of past data suggests extreme caution in predicting what cloud coverage may be.

*The extinction at 0.5 micron is due entirely to scattering, and the numbers are strongly dependent upon the assumed distribution of the size of the water droplets. At 10 microns more than half the extinction is due to absorption and the results are less sensitive to droplet size distribution (see Chapter 4, Section 2).

5.2 Diversity

The fact that even good sites may have substantial (50 percent) cloud cover during several months of the year and nonnegligible cover (10 to 20 percent) over the remaining months indicates that diversity would be required in an optical receiving system. If there are n independent stations, each of which have reliability p, then the over-all reliability is given by

$$P = 1 - (1-p)^n$$

This is illustrated in Table 55 in which the station reliability required to achieve system reliabilities of 90, 95, and 99 percent are shown. For example, three stations that are usable 78 percent of the time would result in a system that is usable 99 percent of the time.

The above oversimplified argument (which unfortunately is frequently used to dismiss the cloud problem) suffers from at least three major deficiencies:

1. Independent statistics are assumed at the separate receiving sites, whereas high-altitude cloud cover and weather patterns tend to be correlated over large areas.

2. The argument discusses the probability of having at least one station not covered by clouds at a given time. It does not discuss the time statistics of the resulting outages. Information on outage statistics are required before any meaningful system design work can be done.

3. The implication of diversity on system operation is not discussed. Many of the implications are obvious and apply to any diversity system; e.g., duplication of facilities with intercommunication (or communication to a central point) required. There are, however, implications peculiar to narrow beam optical systems. These will be discussed below.

The Earth subtends an angle of the order of 10^{-4} radians at Mars distance. A necessary (but by no means sufficient) condition for optical systems to be attractive relative to microwave systems (see Section 6) is the achievement of beamwidths at least an order of magnitude narrower than 10^{-4} radians. For example, a 1-meter diffraction-limited telescope would have a beamwidth of 10^{-5} radians (2 arc seconds) at λ = 10 microns, and a beamwidth of $5(10)^{-7}$ radians (0.1 arc second) at λ = 0.5 microns. At a distance of 10^8 km, the corresponding beam diameters are 10^3 km and 50 km, both well below the diameter of the Earth. Particularly at the visible frequency, diversity stations could not be chosen to be within the beam of the space transmitter and still be able to obtain independent cloud-cover conditions. Consequently, it is necessary for the space vehicle to point to a new station when propagation conditions to a given station deteriorate.

Table 54

BAKER-NUNN SITES SHOWING PERCENT OF TIME LOST DUE TO CLOUDS (TAKEN FROM REFERENCE 3)

Station Coordinates (Long.)	Lat.	Year	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
New Mexico 253°27′ (E)	+32°25′	1962	36	25	27	15	5	12	49	32	48	19	26	22
		1963	34	27	21	29	30	13	48	57	28	16	27	17
South Africa 028°15′ (E)	−25°58′	1962	42	29	31	26	4	6	1	5	8	29	53	37
		1963	50	23	27	28	15	21	20	2	7	30	55	49
Australia 136°46′ (E)	−31°06′	1962	32	21	21	9	34	12	22	22	27	38	15	18
		1963	16	22	22	22	51	35	47	21	18	17	18	22
Spain 353°48′ (E)	+36°28′	1962	52	32	81	43	33	15	21	15	34	47	40	43
		1963	72	49	39	39	35	32	4	19	22	19	60	45
Peru 288°30′ (E)	−16°28′	1962	81	71	56	38	10	2	5	6	35	22	42	55
		1963	88	95	63	56	23	2	6	15	21	26	13	52
Iran 052°31′ (E)	+29°38′	1962	28	44	33	54	10	1	18	16	3	nil	14	38
		1963	17	40	30	47	30	7	15	23	1	12	35	22
Curacao 291°10′ (E)	+12°05′	1962	52	43	56	53	74	66	46	38	60	39	45	50
		1963	70	59	51	68	70	40	64	55	63	50	65	62
Florida 279°53′ (E)	+27°01′	1962	55	40	62	55	36	61	41	55	52	45	55	43
		1963	57	58	47	28	50	42	44	29	49	32	49	43
Argentina 294°54′ (E)	−31°57′	1962	53	52	38	72	51	27	39	32	22	29	35	43
		1963	*	35	40	27	39	33	36	32	57	46	46	30
Hawaii 203°45′ (E)	+20°43′	1962	23	41	56	39	33	5	38	18	20	16	26	33
		1963	59	17	70	78	52	28	29	23	39	39	32	30

*No photography was attempted for a 3-week period when the mirror was removed for realuminizing.

Table 55

DIVERSITY IMPROVEMENT

No. of Stations	Percent Station Reliability for System Reliability of		
	90%	95%	99%
1	90	95	99
2	68	78	90
3	54	63	78
4	44	53	68

This in turn requires a mechanism for recognizing the deterioration of conditions and a mechanism for switching the beam to a new station.

The only reliable method to determine the quality of the optical path between the space vehicle and a given ground station is to measure it. Thus, either the space vehicle requires a broad-beam optical beacon to illuminate the entire Earth or the ground stations require a beacon to illuminate the space vehicle. Since the latter is required in any event for pointing the space-vehicle narrow-communication beam towards a particular earth station (see Chapter 3, Section 4), it is the method which will be discussed here. (Note that if the former, an optical beacon on the spacecraft, were employed, it would still be necessary to communicate the attenuation data back to the spacecraft, although this could presumably be done on a ›wave uplink.)

The beacon acquisition field of view (see Chapter 3, Section 4, and Chapter 4, Section 3) will certainly be wider than the angle subtended by the Earth. If the angular separation of the Earth stations (as viewed from the space vehicle) is greater than the resolution of the acquisition optics, then the separate beacons will give resolved spots on the acquisition photo-detector surface. The resultant photo-current is then proportional to the sum of the beacon powers with no cross terms present. If each beacon is then chopped at a distinct frequency, it is possible to separate the photo-currents by simple filtering and to measure which is largest. Logic could be provided to acquire the strongest such signal and to continue pointing toward that ground station provided the beacon does not drop below a preassigned threshold. (It is not practical to adopt the strategy of always pointing toward the strongest beacon, since this might involve considerable hopping between stations of comparable level.)

Although the above operation is feasible in principle, it does add appreciably to the system complexity. Each Earth beacon would require a distinct chopping code, and the space vehicle must be capable of simultaneously receiving and measuring the level of each such signal. Also, when handing over from one beacon to another, it is necessary to center that beacon in the field of view by the motion of a Risley prism (see Chapter 3, Section 4). During this time, communication is interrupted. The reacquisition is of necessity slow, perhaps of the order of several seconds (see Chapter 3, Section 4 and Chapter 4, Section 3), because of the necessary sluggishness of the Risley prism and the fact that the four-quadrant acquisition system described in Chapter 3, Section 4 gives only the quadrant but not the magnitude of the error until the beam is nearly centered.

Thus, in considering the merits of diversity, it is necessary to consider not only the duplication of ground facilities but also the disadvantages of the more complex spacecraft and the effects of interruption of communication during hand-over. It is necessary to also consider the variability of atmospheric conditions and the possibility for changes during the transit time.

5.3 Atmospheric Fluctuations

Although diversity may counter extended periods of complete attenuation, it cannot counter amplitude fluctuations which are fast compared to the acquisition time but slow compared to the bit period. As noted in Chapter 4, Section 2, such fluctuations may seriously degrade the performance of a digital communication system unless substantial margin (6 dB or more) is provided. As noted in Section 5.2, it is necessary to allow for at least several decibels of atmospheric attenuation before switching to a new station; otherwise, the switching might occur too frequently. Adding "standard" atmospheric attenuation to the above numbers suggests that a ground-based optical system will require at least 12 dB more average signal power than that which would be calculated under free space conditions. This indicates, for example, that a satellite receiver need be only 1/4 the diameter of an earth receiver for the same performance.

Another aspect of the atmospheric fluctuation problem of importance for heterodyne detection at 10 microns* is the lack of spatial coherence over extended distances. Estimates that have been made of correlation distance range from 0.5 to 4 meters (see Chapter 4, Section 2), but actual vertical propagation measurements are required to reliably determine this number. As noted in Chapter 4, Section 6, the effective area of a single heterodyne receiving system is limited by the coherence area of the incoming wavefront.

Larger effective areas may be obtained by using an array of collectors, each smaller than the coherence area. If the heterodyne IF photo-current from each of these collectors is coherently combined, then the resultant effective area is the sum of the area of each of the collectors. However, to perform coherent combining it is necessary to measure the phase of the photo-current at the output of each of the photo-detectors. This may be done if, in addition to the communication sidebands, the received

*As noted in Chapter 4, Sections 6 and 9, there is little reason to consider heterodyne detection at visible frequencies.

signal has a carrier component such that the fraction of the power in the carrier satisfies the inequality

$$\frac{P_c}{P} > \frac{nW_c}{W} \tag{8}$$

where n is the number of signals to be combined, W_c is the bandwidth of the phase-lock loop in which the carrier is recovered and W is the communication bandwidth. Equation (8), which should be interpreted as a functional inequality, rather than a precise numerical inequality, simply states that relative to the communication signal the carrier requires less power because of the reduced bandwidth, but more power because detection must be on the basis of each element of the array. To make carrier recovery feasible, $W_c/W \ll 1$. The bandwidth W_c is limited both by the stability of the laser transmitter and by atmospheric effects. Propagation measurements[4] suggest that the minimum W_c, owing to propagation effects, may be as high as 1 kHz. If the information bandwidth is 1 MHz, a 10-element array could be coherently combined with less than 1 percent of the power devoted to the carrier. However, if one were interested in communication from distances of the order of 10 AU, where smaller communications bandwidth may be of interest (see Section 6), then coherent combining may not be feasible. Further discussion of this appears in Appendix 12.

5.4 Background Noise

As noted in Chapter 4, Section 1, the background noise intensity at optical frequencies, even under day sky conditions, is much less than 1 photon per second per unit bandwidth per unit spatial mode. Consequently, if an optical communication system receives only a single spatial mode and if the noise bandwidth is matched to the communication bandwidth, then background noise may be neglected relative to quantum (signal shot) noise. This situation generally prevails with heterodyne detection. However, it generally does not apply in the case of a direct detection system within the Earth's atmosphere. The remainder of this section will consider the degradation in performance of a 0.5-micron direct-detection system owing to sky noise.

If there is sufficient photomultiplier gain that receiver noise may be neglected, the information rate of a direct detection communication system (see Chapter 4, Section 7) is given by

$$H = \left(\frac{1}{K}\frac{P\eta}{h\nu}\right) \Big/ \left(1 + \frac{N_\nu \Omega W A}{P}\right) \tag{9}$$

where P is the received optical signal power, η is the detector quantum efficiency, K is a constant which depends on the modulation system and the error probability, for binary polarization modulation K = 20 for $P_e = 10^{-5}$, N_ν is the background radiant ter..ity in watts/m^2 -Hz-sr, Ω is the solid angle field of view, is the predetection optical filter bandwidth, and A is the receiver effective area.

Equation (9) may be rewritten in the form

$$\frac{P}{P_o} = \frac{1}{2}\left[1 + \sqrt{1 + 4\alpha}\right] \tag{10}$$

where P_o is the power that would be required in the absence of background noise,

and

$$\alpha \equiv \frac{N_\nu \Omega W A \eta}{K h \nu H} \tag{11}$$

is a parameter which indicates the effect of the background. In Figure 131, P/P_o is plotted as a function of α, both on a dB scale. For $\alpha \gg 1$, $P/P_o \sim \sqrt{\alpha}$.

Consider, for example, $N_\nu = 6(10)^{-14}$ watts/m^2 -Hz-sr, corresponding to a day sky background, $\Omega = 10^{-8}$ sr, $\eta = 0.1$, $W = 10^{11}$ Hz, $A = 50 m^2$, and $H = 10^6$ bits/sec. In this case $\alpha = 40$, which results in an 8-dB degradation (see Figure 131) relative to the noiseless case. Thus, if there is only sufficient power to achieve a communication rate of 10^6 bits/sec, the effect of the background is appreciable. The effect will be even more significant for communication from 10 AU, where smaller information rates are of interest. (This is discussed in more detail in Chapter 4, Section 9.)

A satellite receiver, of course, avoids both the day and night sky background. Narrower fields of view may also be employed to avoid having Mars or other stars within the field of view (see Section 1.2). Thus, background noise may be neglected with a satellite.

5.5 Beacon

As noted in Chapter 4, Section 4, there is a serious problem in achieving sufficient beacon power from a ground transmitter for acquisition in the spacecraft. The problem is more serious for a ground as opposed to a satellite beacon because:

1. The minimum beacon beamwidth is limited by atmospheric effects to about 10^{-4} radians.
2. The beacon must be received in the presence of earth shine.

Although a satellite beacon could in principle be narrowed below 10^{-4} radian, in practice the necessity of having the beacon illuminate the spacecraft on an open-loop basis argues against this. Considerable advantage is obtained, however, by having the beacon located so that the Earth is not within the field of view of the acquisition receiver. If the beacon is on an Earth-synchronous satellite and if the spacecraft is in the vicinity of Mars and has an acquisition receiver with a beamwidth of 10^{-4} radian, then, under worst-case conditions, for 10 percent of the synchronous-satellite orbit, the Earth will be within the field of view. Because of the 23-degree angle between the plane of the

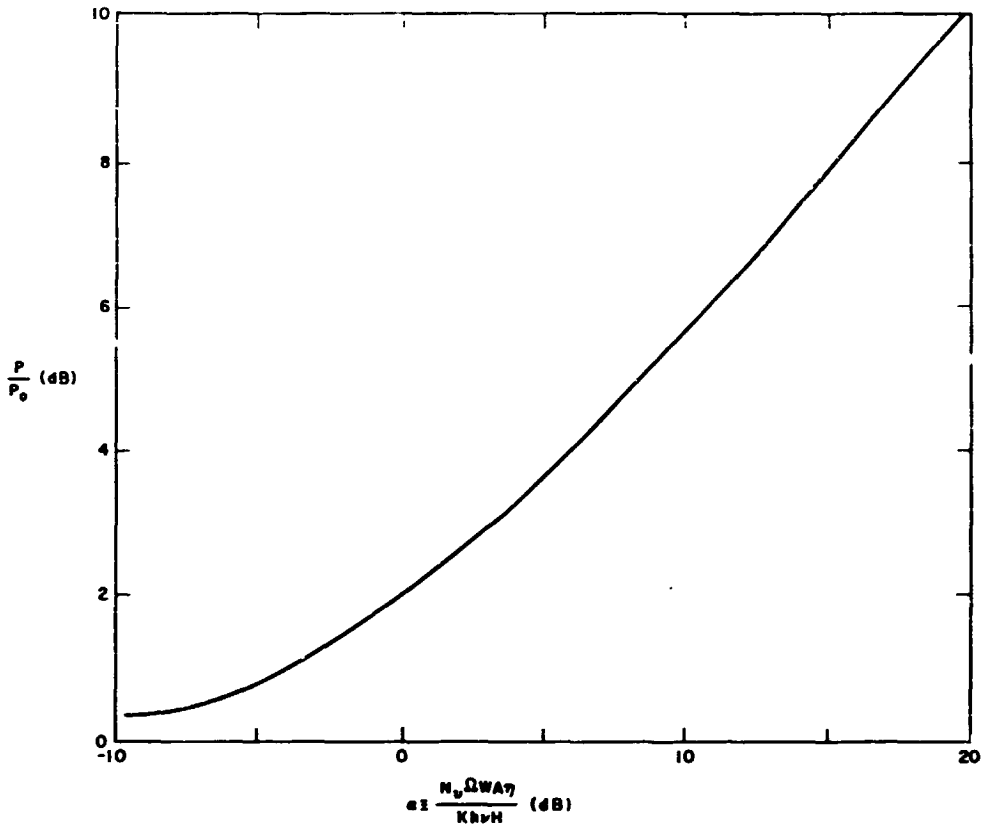


Figure 131. Increase in transmitter power necessitated by background noise

ecliptic and the Earth equatorial plane, there will be extended periods during which the Earth will not be within the field of view (see Section 1). In the absence of Earth shine, a CW beacon power of 3 watts would suffice as compared to 200 watts from an Earth beacon (see Figure 102). Thus the satellite receiver permits a beacon with more reasonable power requirements. It should be noted, however, that the above numbers correspond to a spacecraft in the vicinity of Mars. If the distance is an order of magnitude greater, then the beacon acquisition appears impossible independent of whether the beacon is on the Earth or in a satellite.

5.6 Hand-Over

The same logic that is used in diversity switching (Section 5.2) may also be employed to effect hand-over when a ground terminal is no longer visible to the spacecraft. As the beacon from a given ground station fades out due to shadowing*, the spacecraft acquisition system automatically acquires the strongest beacon within the field of view. However, as described in Section 5.2, this results in an interruption in communication unless multiple beams are employed with the spacecraft transmitter. Although this latter alternative has not been investigated in any detail, it does not appear practical, particularly when the vagaries of atmospheric transmission are considered. Thus, in the case of ground reception, even in the absence of atmospheric attenuation, communication will be interrupted at least twice per day.

As noted in Section 1, a Mars-synchronous satellite will be continuously visible from an Earth-synchronous satellite for periods of several months. This is a particularly compelling argument for the use of a satellite receiver, since only a single satellite and a single ground station need be employed, with no concern for the problems of weather or hand-over.

For all of the above reasons, the calculation of optical system performance in the following section will be for a satellite receiving system. Some comparisons will be made, however, with ground-based receivers. The latter may be of interest, for example, in applications in which occasional loss of the high-speed data is acceptable.

6. PERFORMANCE OF OPTICAL SYSTEMS

In Chapter 4, Section 9, various optical communication systems were compared, and it was concluded that at visible frequencies direct detection is a clear choice over heterodyne or optical preamplifier systems, whereas in the infrared (10 micron) direct detection is far inferior to either heterodyne or optical preamplifier systems. The idealized performance calculations indicate that megabit communication rates from Mars distance are feasible, and the infrared systems appear to have the edge

This section evaluates the performance of optical communication systems relative to the microwave and millimeter performance evaluated in Sections 2 and 3. Optical systems permit a transmitting antenna gain that is several orders of magnitude greater than that achievable at microwave or millimeter frequencies. However, all of the other factors in the transmission equation [Equation (7), Section 2.1]: viz., transmitter power, receiver aperture, noise temperature, and modulation power efficiency, tend to be poorer at the optical frequencies.

6.1 Transmitter Power

Considering efficiency, high power capability, and detector performance, the leading laser candidate in the visible region is Nd:YAG with second harmonic generation ($\lambda = 0.53$ microns). As indicated in Chapter 3, Section 1, single-mode power of 6 watts has been obtained with an overall efficiency† of 0.2 percent. Both the efficiency and the lifetime are limited by the pump lamp. With better crystals for second harmonic generation an efficiency of 0.3 percent should be achievable with present lamps; substantial further improvement is possible with new types of lamps that radiate more energy in the pumping band.

By far the best efficiencies and highest power have been achieved in the infrared with the CO_2 laser (10.6 microns). For a flowing gas system, single mode power of 10 watts has been achieved with an efficiency of 8 percent, although higher efficiencies have been achieved under multi-mode and/or low power conditions.

Thus the efficiency of lasers is considerably smaller than that of microwave and millimeter wave tubes. This, coupled with the fact that laser efficiency generally diminishes with increased operating temperature, poses a more severe thermal problem for lasers than for microwave power amplifiers.

In Table 56 the laser power output, relative to a 40 percent efficiency microwave system, is shown for two cases: (1) constant prime power and (2) constant dissipated power. The thermal problem could be controlling, so the second column in Table 56 might be more significant.

Thus, compared to a microwave system, a laser system may be expected to radiate 9 dB less power at 10.6 microns and 23 dB less power at 0.53 microns. For example, if the maximum power that may be dissipated is 300 watts, then the microwave system would radiate 200 watts, the 10.6 micron system would radiate 25 watts, and the 0.53 micron system would radiate 1 watt.

*In practice it would be better to turn off the beacon when the spacecraft is below, say, 10 degrees elevation.

†The efficiency considered here is the ratio of optical output power to low-voltage dc primary power (but excluding power required for cooling).

Table 56

LASER OUTPUT POWER RELATIVE TO 40 PERCENT EFFICIENT MICROWAVE POWER
(Prime Power, 1-5 kW Range)

	Constant Prime Power (dB)	Constant Dissipated Power (dB)
0.53μ (0.3% efficiency)	-21	-23
10.6μ (8% efficiency)	- 7	- 9

It may be argued that the efficiency of subsequent laser systems have more room for improvement than microwave systems and, consequently, that the above differences will diminish with time. On the other hand, the optical system has been given the benefit of the doubt that the same heat may be dissipated as in a microwave system, and in assuming that comparable reliability may be achieved. Also, no allowances have been made for losses in the optical modulator.

6.2 Telescope Gain

The gain of the transmitting telescope is determined by the size of the objective mirror and the wavelength. Assuming a 70 percent aperture efficiency, the gain is given by:

$$G = 7(D/\lambda)^2 \qquad (12)$$

Thus for D = 1 meter; G = 134 dB at $\lambda = 0.53\mu$, and G = 108 dB at $\lambda = 10.6\mu$. At microwave frequencies, spacecraft antenna gain is generally limited (see Section 2) to 50 to 60 dB, so that optical frequencies offer the possibility for substantial improvement.

Unfortunately there is rather limited information concerning the weight of space telescopes. In Figure 132, weight (see Table 32, Chapter 3, Section 3) is plotted as a function of diameter. It is seen that there is a prohibitive increase in weight (associated with active, segmented optics) for telescope diameters in excess of 1 meter. These results are for telescopes which are diffraction limited in the visible region. At 10.6 microns it may be possible to build larger telescopes with a smaller weight penalty, but specific designs of 10.6 micron space telescopes do not exist. It is reasonable to assume, however, that extension of optical techniques should permit construction of diffraction limited telescopes of perhaps twice the aperture area at 10.6μ at the same total mass.

In Figure 133, the gain of microwave (2 and 8 GHz) spacecraft antennas, and the gain of an optical telescope (at $\lambda = 0.53$ micron) are shown as a function of weight. It is seen that the gain advantage of the optical telescope increases as the weight increases. This opposes the trend that was observed at microwave and millimeter frequencies where the higher frequencies are generally more advantageous at low weight. (This is also apparent in Figure 133 where the curves at 2 GHz and 8 GHz are further apart at low weight than at high weight.)

It follows from Figure 133 that for a telescope and antenna weight of 250 pounds, the optical system achieves a gain advantage of 75 dB with respect to a 2 GHz system, and 66 dB with respect to an 8 GHz system. For a telescope and antenna weight of 1000 pounds, the corresponding advantages are 83 dB and 75 dB.

6.3 Receiving Effective Area

In the case of an Earth-based telescope, cost is an appropriate parameter to compare with the cost of large receiving microwave antennas. Thus the cost of the 210 foot Goldstone antenna is about the same as that of the 200 inch Palomar telescope (See Chapter 1, Section 5 and Chapter 3, Section 5)*. However, as noted in Section 5, there are compelling reasons for employing a satellite-based rather than an Earth-based receiving telescope. In this case cost is difficult to assess – it includes not only the cost of the telescope, but also launch costs, and the costs associated with the microwave link from the satellite to the earth. Furthermore, the effective cost is strongly influenced by reliability and lifetime considerations.

The approach taken here will be to consider a synchronous-satellite receiving telescope of 1.4 meter diameter, consistent with the 10.6μ transmitting antenna. If this diameter is indeed feasible for the deep space probe, it should be feasible for an Earth satellite of similar cost.† The same diameter is assumed for both 0.53μ and 10.6μ since the telescope need not be diffraction limited in the visible. It would need, however, to be diffraction limited at 10.6 microns for heterodyne detection systems.

Compared with a 64 meter (210 foot) microwave receiving antenna, the 1.4 meter receiving telescope has 33 dB less receiving effective area.

6.4 Noise Temperature

For a satellite receiving system, background noise may be neglected. Also, with the use of photomultiplier direct detection at 0.5 micron and heterodyne detection at 10.6

*As noted in Chapters 1 and 3, there is considerable variability in the cost of microwave antennas and optical telescopes. As a rough rule of thumb, costs are about the same for a telescope 1/12 the diameter of a steerable microwave antenna. Thus, both an 80 inch telescope and an 80 foot S-band antenna cost about $1 million. The corresponding cost for 200 inches and 200 feet is about $12 million.

†There is a possible loophole in this argument because the thermal environment of an Earth satellite is more severe than that of a deep space probe.

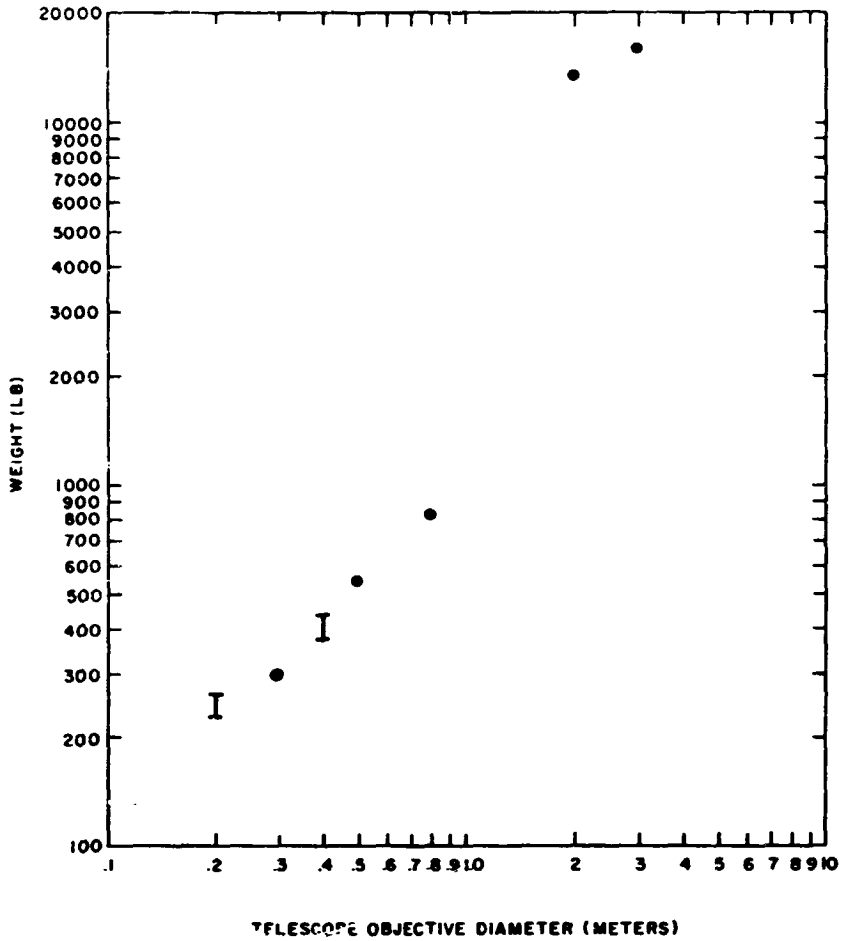


Figure 132. Weight of space telescopes

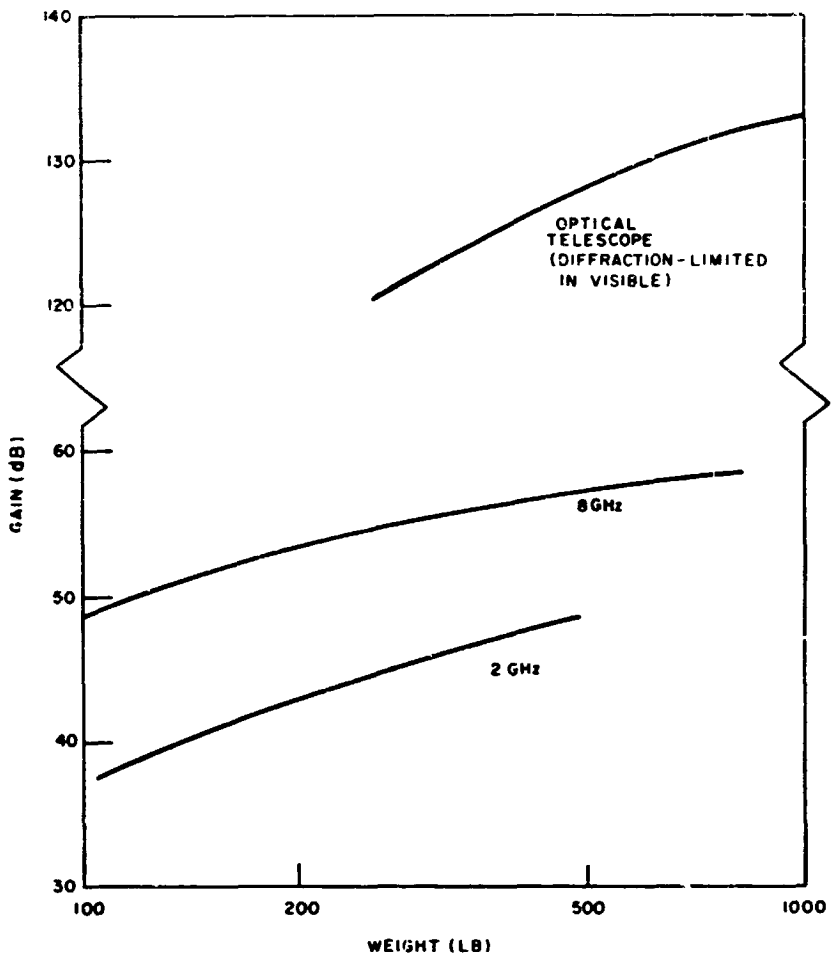


Figure 133. Comparison of the gain of space telescopes and microwave antennas

microns, the current at the output of the detector is sufficiently large (see Chapter 4, Sections 6 and 7) that detector dark current and Johnson noise may be neglected. In this case the dominant noise term is signal shot-noise (also called quantum noise). For heterodyne detection (Chapter 4, Section 6) this noise is equivalent to additive Gaussian noise with spectral density $N_o = h\nu$, where $h = 6.634(10)^{-34}$ joules-sec is Planck's constant, and ν is the center frequency. The minimum equivalent system noise temperature is then given by

$$T = h\nu/k \qquad (13)$$

where $k = 1.386(10)^{-23}$ joule/°K is Boltzmann's constant. Equation (13) also gives the minimum noise temperature of an ideal laser amplifier (Chapter 4, Section 8).

Although in the case of direct detection, the shot noise cannot be rigorously treated as additive noise of spectral density $h\nu$, as shown in Chapter 4, Section 7, this is an adequate approximation for most cases of practical interest.

It follows from Equation (13) that the minimum noise temperature is 1350°K at $\lambda = 10.6$ microns and 27,000°K at $\lambda = 0.53$ micron. Compared to a 25°K noise temperature at 2 GHz, there is a 17 dB noise penalty at 10.6 microns and a 30 dB penalty at 0.53 micron.

6.5 Losses

By employing a satellite receiver, atmospheric losses need not be considered. However, there are several other sources of loss in optical receivers. Direct-detection systems in the visible and heterodyne systems in the infrared are considered below.

Although optical preamplifiers alleviate some of these losses, as indicated in Chapter 4, Section 9, the preamplifier does not offer an appreciable net advantage at 0.53 micron. The dominant loss is the quantum efficiency of the detector. As noted in Chapter 3, Section 7.1, selected photo-emissive surfaces are available with quantum efficiencies of 0.20 at 0.53 micron corresponding a loss of 7 dB. In addition to the detector loss, there will be losses in the optical filter and the polarizer (assuming polarization modulation as described in Chapter 4, Section 7). However, since background is not a serious problem with a satellite receiver, there is no need to employ an unusually narrow filter, and with careful design, it should be possible to keep optical losses to no more than 3 dB.

In the case of heterodyne detection at 10.6 microns, the quantum efficiency of photoconductors (which, however, require operating temperatures below 77°K; see chapter 3, Section 7) is as high as 50 percent corresponding to a 3 dB loss. There is not the need for an optical filter or polarizer. Although a beam splitter is required to combine the signal and local oscillator fields, there need be little attenuation in the signal path since a local oscillator power of the order of 1 milliwatt (measured at the detector) is sufficient to obtain the quantum noise limit (see Chapter 4, Section 6). Thus, in the optical combining, the signal attenuation can be kept well below 1 dB by use of a higher power of local oscillator (e.g., a high reflectivity mirror may be used for reflection of the signal and transmission of the local oscillator).

There is, however, an effective signal loss in heterodyne systems due to imperfect matching (both amplitude distribution and alignment; see Chapter 4, Section 5) of the signal and local oscillator spots on the detector.* It will be assumed, rather optimistically, that the total optical and alignment loss can be kept to 1 dB.

6.6 Communication Performance

The parameter, E/N_o, used to describe the power efficiency of digital modulation systems takes on a simple interpretation at optical frequencies. Since E is the energy per bit and $N_o = h\nu$ is the energy per photon, E/N_o is the number of photons required per bit.

In principle, the coherent modulation and coding techniques that are employed at microwave frequencies may also be employed with a heterodyne detection laser system. Consequently, it will be assumed that heterodyne systems may operate at the same E/N_o as coherent microwave systems; viz., $E/N_o \approx 5$ dB, as obtained with biorthogonal modulation.

In the case of direct detection optical systems, the situation is different in two respects. First, direct detection systems necessitate incoherent modulation which generally requires larger E/N_o than do coherent techniques. On the other hand, a direct detection system which operates at the shot noise limit (i.e., both background and dark current negligibly small) has noise present only when the signal is present. This contrasts with a heterodyne system in which a continuous local oscillator implies a continuous source of shot noise. Analysis indicates that both direct detection binary polarization modulation and heterodyne detection binary phase shift keying achieve essentially the same performance; viz., $E/N_o \approx 10$ dB for $P_e = 10^{-5}$.

As noted above, appropriate coding of the coherent system (biorthogonal modulation), can result in about a 5 dB reduction in E/N_o. Although coding may also be applied in the direct detection case, one is generally constrained to decoding techniques in which binary decisions are made on the basis of the individual pulses, and these binary decisions are processed algebraically in the decoder. This binary quantization, prior to the decoder, results typically in a 2 dB penalty relative to what may be achieved in the coherent case.

*Indeed, to provide for continuous tracking and to more simply generate a strong error signal, it is desirable to have the local oscillator spot several times the diameter of the signal spot.

On the basis of the above considerations it will be assumed that the heterodyne systems may operate at the same E/N_0 as coherent microwave systems ($E/N_0 \approx 3$ dB), but that direct detection incoherent systems suffer a 2 dB penalty ($E/N_0 \approx 7$ dB).

Consider next the question of possible future growth. Sequential decoding offers the promise of achieving $E/N_0 = 3$ dB with both the microwave and heterodyne optical systems. In direct detection optical systems, pulse position modulation, with a large number of pulse positions, offers the greatest promise for a reduction in E/N_0. Consider, for example, a PPM system with a basic frame period of 10 microseconds; each frame being divided into 2^{10} time slots of duration 10 nanoseconds each. By transmitting a pulse in one of these 2^{10} slots, each pulse conveys 10 bits of information, and the information rate is 10^6 bits per second. In the shot noise limit, 10 photons are required per pulse which leads to an E/N_0 of 0 dB. Although such systems are considered in New Technology Chapter and Appendix 9, they are judged presently impractical because of the conflicting requirements of high average power, high average prf, and variable interpulse times. Mode dumping techniques are generally restricted to kilohertz pulse repetition rates, so that a two-order-of-magnitude improvement would be required to achieve the above performance.

On the basis of the above considerations, it will be assumed that future growth may allow a 2 dB reduction in E/N_0 for the microwave and coherent optical systems (viz. $E/N_0 \approx 3$ dB), and up to a 7 dB reduction in the E/N_0 for the incoherent optical systems (viz. $E/N_0 \approx 0$ dB). Thus, in terms of relative performance, the incoherent systems are judged to have a 2 dB poorer E/N_0 than the coherent systems, with a possible growth potential that would convert this to a 3 dB advantage.

Table 57 summarizes performance of the 10.6 micron and 0.53 micron systems relative to a 2.3 gHz system. The assumptions used in Table 57 are listed in Table 58.

With these assumptions, the infrared heterodyne system achieves essentially the same performance as the microwave system, whereas the visible system is more limited. It can be concluded that optical communication systems achieve high capacity communication from deep space, but microwave systems can achieve similar or better performance more readily. It should also be noted that, although the 10.6 micron system is (according to Table 57) 12 dB better than the 0.53 micron system, this presumes the development of appropriate tracking techniques for heterodyne receivers, highspeed detectors, and materials capable of providing modulation of high power signals with low loss and low power consumption. Also not reflected in these numbers are the more difficult problems of acquisition and tracking necessary to establish and maintain the optical communication links, as discussed in Chapter 3, Section 4, and Chapter 4, Section 3. It may not even be possible to maintain lock of an optical transmitter at distances greater than 1 AU.

Table 57

PERFORMANCE OF OPTICAL SYSTEMS RELATIVE TO 2 GHz MICROWAVE SYSTEMS
(All Entries in Decibels)

	10.6 microns heterodyne	0.53 micron direct detection
Power	- 7	-21
Transmitting gain	+60	+83
Receiving area	-33	-33
Noise temperature	-17	-30
Receiver losses	- 4	-10
E/N_0	0	- 2
Net advantage	- 1	-13

Table 8B

ASSUMPTIONS USED FOR TABLE 8-7

	2.3 GHz	10.6 microns	0.53 micron
Efficiency (%)	40	8	0.3
Transmitting gain (dB) (Antenna and optics)	81 (Diameter 7.2 m)	111 (Diameter 1.4m)	134 (Diameter=1m)
Receiving diameter (meters) Location	64 Earth	1.4 Relay satellite	1.4 Relay satellite
Noise temperature (°K)	25	1,380 (Quantum limit)	27,000 (Quantum limit)
Receiver losses	≪1 dB		
Quantum efficiency		0.5	0.2
Overall optical efficiency		0.8	0.5
Modulation	Coherent biorthogonal	Coherent biorthogonal	Incoherent polarization shift

REFERENCES

1. P.D. Potter, W.D. Merrick, and A.C. Ludwig, "Big Antenna Systems for Deep Space Communications," Astronautics and Aeronautics (October 1966), pp 84-95.
2. Study on Optical Communication Experimental Facility, Final Report, Sylvania Electronic Systems, Waltham, Massachusetts, Contract NAS8-20304, 1967.
3. F. Kalil, Optical and Microwave Communications – A Comparison, Goddard Space Flight Center, Report X-507-66-173, March 1966.
4. J.A. Collinson and M. Subramanian, "Modulation of Laser Beams by Atmospheric Turbulence – Depth of Modulation," BSTJ, 46 (March 1967), p 623.

CHAPTER 6.

TRACKING AND NAVIGATION STUDIES

1. INTRODUCTION

The tracking and navigation studies described in this chapter are a necessary adjunct to the communication system studies presented in the preceding chapters. In particular, the navigation function provides the line of sight at all times for communication. For beamwidths in the rf category this clearly does not present any difficulty, but at optical frequencies it becomes a more interesting problem.

The philosophy adopted in the numerical studies was to use the simplest possible geometric and mathematical models of the tracking process by which the essential system tradeoffs could be illuminated. In some cases the "first-cut" calculations were superceded by more elaborate formulations after the essential features had been identified with the simpler model.

Since a numerical simulation has to address itself to a particular mission profile, an Earth-Mars flyby and an intragalactic mission were chosen as typical object studies. The former includes a Mars orbiter as a navigation aid for the passing space probe. However, no attempt was made to simulate the detailed visibility history in any specific case, since variations in coverage "average out" for navigational purposes over a long mission. Since the process of error propagation as a function of various mission profiles is too complex to be expressed in a simple analytic form for parametric trade-offs, one resorts to an ensemble of individual case studies in order to draw general conclusions. These mainly concern the relative effectiveness of different kinds of observational data and tracking modes. The following examples are a step in that direction; particular emphasis is placed on comparisons between rf and optical tracking.

1.1 Navigation for Transfer Trajectories to Mars

Section 2 concerns steady-state navigation during transfer from Earth to Mars. This error study relies on a digital algorithm for refining upon the state covariance matrix of the vehicle according to a maximum-likelihood processor. It allows for uncorrelated random errors in the measurements and bias errors, where the latter must be added to the state vector in carrying the analysis across midcourse corrections. The updating of the covariance matrix to successively later times in the mission is based on a digital integration of the differential equation for the transition matrix.

The numerical studies compare navigational accuracies attainable with r, $\dot{r}$ data from rf tracking alone and with the inclusion of optical angle measurements. In assuming realistic accuracies for the tracking data, serving as input for the orbit refinement, some judgment had to be exercised. While tracking precisions such as $\sigma_r = 10$ m and $\sigma_{\dot{r}} = 1$-3 mm/sec are typical for DSIF-type equipment (see, for example, Reference 5), the accuracies of range and range rate from the trajectory determination will be less than that, in view of geophysical uncertainties -- in particular, the probable error in the speed of light. Strictly speaking, these effects should have been included as bias errors in the statistical model but, for the sake of simplicity, this was not done. Instead, a range of input characteristics was used, with σ_r varying from 100 m to 100 km and $\sigma_{\dot{r}}$ from 1 mm/s to 3 mm/s. The largest standard deviations were taken as a conservative representation of the current state of the art while the high precision data were considered as projections to the future when geophysical biases are better in hand. In this simplified treatment, we have also neglected the fact that σ_r and $\sigma_{\dot{r}}$ are functions of range and vary over the course of a typical interplanetary mission as the range goes from fractions of an astronomical unit to several astronomical units.*

For the optics a σ of 15 or 20 s was taken to represent angle measurements through the atmosphere, while $\sigma = 2$ s or better was representative of star trackers and astrometric measurements. The contribution from high-quality optics, if added to conventional r, $\dot{r}$ data, is most noticeable during the near-Earth phase of the mission, together with beneficial effects from changes in the tracking geometry and accumulation of data. After midcourse corrections the advantages of high-grade optics are again noticeable in

*See, for example, Interplanetary Midcourse Guidance Using Radar Tracking and On-Board Observation Data, by L. S. Cicolani, NASA TN D-3623.

terms of accelerated orbit refinement. However, these payoffs must be expected to decrease in the future as smaller values for σ_r and $\sigma_{\dot{r}}$ become realistic.

A drastic difference is observed between the error propagation with a weighted least-square processor (as used in several past space missions) and an optimal processor. Indeed, some cases show an increase in the rms position error with the former, in spite of data accumulation, whereas the latter causes an effective suppression of errors, as expected. (The eventual velocity error buildup after flight times of about 200 days reflects the growing perturbations due to Mars.)

Several improvements of this error-propagation model are possible but have been omitted as insignificant to the present system studies. Thus, for example, the biases should include a realistic representation of tracker location errors, some of which are to be placed in synchronous orbits around the Earth. Also, the uncertainties in various astrophysical constants should be accounted for. Here some rather exotic phenomena, such as vagaries of the Earth's rotation, wandering of the poles, and relativistic effects, will influence the absolute accuracy of orbit prediction for missions of long duration.

1.2 Navigation Near Earth

An important aspect of most navigation schemes is the tracking that takes place near Earth, this being the place for correction of errors incurred during the injection maneuver. The covariance matrix resulting from this phase represents the initial estimate supplied to the long-range error propagation discussed in Section 1.1.

At short distances from Earth one may enhance the navigational accuracy significantly by the trilateration and triangulation schemes discussed in Section 3 and depicted in Figure 143. A pair of tracking relays is used, at the stable Earth-Moon libration points* or in synchronous orbits, to yield the necessary base line. In order to minimize computational complexities, a two-dimensional model was treated, with the trackers in a circular orbit, the space probe trajectory approximated by a straight line, and the orbit refinement represented as a continuous process. The resulting error histories show a sharp initial decline from the wide baseline and data accumulation. As would be expected, the addition of angle data significantly enhances the estimates of trajectory parameters. The rms errors of the input data were chosen conservatively for the same reasons as in the preceding study, i.e., to allow for geophysical uncertainties. As optical measurements are refined beyond 1 s of arc, it turns out that such measurements from near-Earth satellites can eliminate the need for rf trilateration or triangulation from stations at the libration points.

This two-dimensional model is thought to cover all the salient points for a system study. To be sure, the simulation could be generalized by going to three dimensions and including a more elaborate model for the position errors of the tracking relays. Indeed, the latter should be envisioned as output from a separate orbit determination for these satellites and could lead to very interesting ramifications. However, since these details will probably not affect the overall tradeoffs considered here, they were disregarded for the present.

1.3 Terminal Navigation Near Mars

The error study of navigation during a Mars flyby is in many respects an analog of the near-Earth situation. In Section 4 a similar tracking relay is assumed in an elliptic orbit about Mars and the space probe is traveling on a hyperbolic trajectory. Again, a two-dimensional representation of the encounter is expected to contain most salient features of the tracking problem as shown in Figure 145. The data processing is modeled in an intermittent fashion and based on r, $\dot{r}$ data for the probe and Mars orbiter, each taken from near-Earth trackers as well as relative to each other. The covariance matrix is updated analytically.

One significant difference from the near-Earth simulation is that the state vector of the observation satellite was subjected to refinement, together with that of the spaceship. Both would be tracked from Earth during the long transit phase of the probe from Earth to Mars. The orbit of the satellite becomes quite well established in that time. During the critical hours of the flyby past the planet, the orbiter serves as a nearby reference for the spacecraft. This constitutes one of the interesting features in this investigation. In particular, the recovery of spacecraft ephemeris accuracy after transient disturbances is greatly expedited with the help of a Mars orbiter over Earth-based tracking alone. Optical angle data were not found to affect this contrast between Earth-based and satellite-based tracking in any significant way and were not included in the plots of results.

As in the other studies, obvious generalizations consist of including the third dimension, treating biases properly, and adding other types of observations, such as three-way Doppler measurements between Earth, space vehicle, and orbiter, as well as intermittent astrometric readings from the spacecraft. In view of the great importance of on-board operations in most flyby missions, some of these extensions may carry more weight from the systems point of view than the ramifications listed in Sections 1.1 and 1.2.

*We restrict ourselves to the points commonly known as L_4 and L_5, which lie on either side of the Earth-Moon line and form a quadrilateral with the two primary bodies. In the rotating framework of this quadrilateral, L_4 and L_5 represent stable equilibrium positions for orbiting vehicles, as shown in the theory of the restricted three-body problem.

1.4 Intragalactic Navigation

This section gives an error analysis of the steady-state navigation process for trajectories leading out of the solar system. It parallels the Earth-Mars transfer studies in concept and methodology. All important phenomena, such as error decay due to changing geometry, data accumulation, and the use of optical angle measurements, repeat themselves at a different geometric scale. Indeed, the use of exoatmospheric optics together with DSIF-grade r, $\dot{r}$ data shows its usual beneficial effect. As in the interplanetary case, the ephemeris gradually deteriorates as the tracking distances become very large.

The above remarks summarize what is described more fully in the remainder of this chapter. While an effort was made to conduct these navigation studies as background for the various system comparisons for earlier chapters, a certain looseness of coupling between the work in statistical navigation and communication engineering is undeniable. This may change for manned missions, where abort schemes and autonomous navigation methods are essential parts of any realistic system concept. In such cases the balance between the tracking and communication modes of operation for a given set of rf or optical equipment can change noticeably in that the navigation function can place high-priority, though intermittent, demands on the communication capability of the system. These considerations are particularly germane to follow-on items connected with terminal navigation studies (Section 4).

2. NAVIGATION FOR TRANSFER TRAJECTORIES TO MARS

This section discusses the accuracy with which the orbit of an interplanetary probe can be determined for different assumptions about the accuracy and type of available measurements. Specifically, orbit determination with range and range rate measurements is studied both with and without optical inputs; i.e., highly accurate azimuth and elevation angles. Random and bias errors in the measurements are also considered. As actual in-flight data processing is frequently performed using weighted least squares, the accuracies obtained with this procedure are analyzed. However, since improved accuracies can be obtained using optimal data processing, this case is analyzed as well. Finally, the section covers the effect of the presence of optical inputs on the accuracies obtainable after a midcourse correction.

The computer program used in the above studies was a modification and extension of one described elsewhere.[1] The program computes the trajectory of the probe and the transition matrices by numerical integration, taking into account the gravitational attractions of Sun, Earth, and Mars. Some details of the expressions used in the analysis of the data processing are given in the next several paragraphs.

2.1 Method of Analysis

The following formula[1] gives the covariance matrix of the errors in the position and velocity components of the probe (μ, a 1 X 6 column matrix) after processing of m observations (λ, a 1 X m column matrix)

$$A_c = \mathrm{cov}(\delta\mu) = (J^T W J)^{-1} J^T W \left[N \Lambda_b N^T + \Lambda_r \right] W J (J^T W J)^{-1} \qquad (1)$$

The undefined quantities in Equation (1) are

$J = \dfrac{\partial \lambda}{\partial \mu}$, a m X 6 matrix

$N = \dfrac{\partial \lambda}{\partial \nu}$, a m X b matrix of bias sensitivities, where ν is a 1 X b matrix of bias sources

Λ_r , a m X m diagonal covariance matrix of the random errors

Λ_b , a b X b diagonal covariance matrix of the bias errors

W , a m X m weighting matrix used in the data processing.

A detailed derivation of Equation (1) is given in Reference 1, pp. B-4 to B-6, and will not be repeated here. This discussion will be limited to certain variations pertinent to this study, and also to certain computational details which are the difference between merely having a correct formula and also obtaining correct results.

The matrix W is arbitrary in Equation (1), although the equation was derived to investigate the case of least-squares data processing, where W is a diagonal matrix. The case where W is the optimal weighting matrix will be considered because significant improvements in orbit determination accuracy result when this is implemented.

Assume that the errors in the observations are of the form

$$\epsilon = \epsilon' + N\epsilon''$$

where the components of ϵ' are m sample values of m statistically independent random variables with zero means and covariance matrix Λ_r and the b components of ϵ'' are sample values of b random variables with zero means and covariance matrix Λ_b. Then the optimal choice of W is given by[2]

$$W = \left(\Lambda_r + N \Lambda_b N^T \right)^{-1} \qquad (2)$$

Equation (2) for W (a m X m matrix, in general non-diagonal) can be inverted analytically by use of the matrix inversion lemma[2] to yield

$$W = \Lambda_r^{-1} - \Lambda_r^{-1} N \left(\Lambda_b^{-1} + N^T \Lambda_r^{-1} N \right)^{-1} N^T \Lambda_r^{-1} \qquad (3)$$

Insertion of Equation (2) into Equation (1) yields

$$A_t = \left(J^T W J\right)^{-1} \tag{4}$$

with W given in Equation (3). Equation (4), with $\Lambda_b = 0$, is the well-known result for optimal processing in the presence of purely random errors, where $W = \Lambda_r^{-1}$. Equations (3) and (4) yield an explicit expression for the inverse covariance matrix

$$A_t^{-1} = J^T \Lambda_r^{-1} J - J^T \Lambda_r^{-1} N \left(\Lambda_b^{-1} + N^T \Lambda_r^{-1} N\right)^{-1} \left(J^T \Lambda_r^{-1} N\right)^T \tag{5}$$

Use of Equation (5) directly would be extremely unwise computationally because the multiplication of matrices of high dimension will be called for if m is a large number. Equation (5) must be put in a suitable form for so-called "batch processing."* Assume that the m observations are composed of $m_r, m_{\dot r}, m_A, m_E$ range, range rate, azimuth, or elevation measurements where

$$\sum_i m_i = m \qquad i = r, \dot r, A, E$$

and any of the m_i may equal zero. Then, upon suitable partitioning of the matrices appearing in Equation (5),

$$A_t^{-1} = \sum_i \left\{ J_i^T \Lambda_{r_i}^{-1} J_i - J_i^T \Lambda_{r_i}^{-1} N_i \left(\Lambda_{b_i}^{-1} + N_i^T \Lambda_{r_i}^{-1} N_i\right)^{-1} \left(J_i^T \Lambda_{r_i}^{-1} N_i\right)^T \right\} \tag{6}$$

where $J_i = \dfrac{\partial \lambda_i}{\partial \mu}$ a $m_i \times 6$ matrix, with λ_i the $1 \times m_i$ column of measurements of type i

$N_i = \dfrac{\partial \lambda_i}{\partial \nu}$, a $m_i \times 1$ matrix†

$\Lambda_{r_i} = \sigma_i^2 I$, with I the $m_i \times m_i$ unit matrix and σ_i^2 the variance of the i^{th} type of measurement

$\Lambda_{b_i} = \gamma_i^2$, the variance of the bias in the i^{th} type of measurement

Equation (6) consist of four terms (one for each type of measurement) each of the same form as Equation (5). These data must be processed sequentially, thus computing A_t as successive batches of data are processed. If the component matrices of Equation (6) are partitioned properly and simplified

$$A_t^{-1} = \sum_i \left\{ \sigma_i^{-2} J_i^T J_i - \sigma_i^{-4} w_i^{-1} J_i^T N_i \left(J_i^T N_i\right)^T \right\}$$

$$i = r, \dot r, A, E \tag{7}$$

where

$$J_i^T J_i = \sum_{m_i} \begin{bmatrix} \dfrac{\partial \lambda_{m_i}}{\partial \mu_1} \\ \cdots \\ \dfrac{\partial \lambda_{m_i}}{\partial \mu_6} \end{bmatrix} \begin{bmatrix} \dfrac{\partial \lambda_{m_i}}{\partial \mu_1} & \cdots & \dfrac{\partial \lambda_{m_i}}{\partial \mu_6} \end{bmatrix}$$

$$J_i^T N_i = \begin{bmatrix} \displaystyle\sum_{m_i} \dfrac{\partial \lambda_{m_i}}{\partial \mu_1} \\ \cdots \\ \displaystyle\sum_{m_i} \dfrac{\partial \lambda_{m_i}}{\partial \mu_6} \end{bmatrix}$$

$$w_i^{-1} = \gamma_i^2 \left\{ 1 + m_i \left(\gamma_i^2 / \sigma_i^2\right) \right\}^{-1}$$

The expression w_i^{-1} has been written in a form which is determinate even if there are no bias errors ($\gamma_i = 0$) so that Equation (7) can be used to analyze optimal processing of purely random or random and bias errors.

For subsequent incorporation of the effect of midcourse corrections, Equation (7) can be rewritten in the form

$$A_t^{-1} = Q - P Z^{-1} P^T \tag{8}$$

where

$$Q = \sum_i \sum_{m_i} \sigma_i^{-2} \left(\frac{\partial \lambda_{m_i}}{\partial \mu}\right)^T \left(\frac{\partial \lambda_{m_i}}{\partial \mu}\right) \quad (6 \times 6)$$

*An analogous procedure for Equation (1) is described in Reference 1 and in Section 2.2 of this chapter.

†For simplicity it is assumed that bias errors occur directly in the measurements only (i.e., not station location errors, errors in physical constants, etc.). Thus N_i is a column of m_i 1's.

$$P = \sum_{m_i} \left[\sigma_r^{-2} \frac{\partial r}{\partial \mu} \;\vdots\; \sigma_{\dot{r}}^{-2} \frac{\partial \dot{r}}{\partial \mu} \;\vdots\; \sigma_A^{-2} \frac{\partial A}{\partial \mu} \;\vdots\; \sigma_E^{-2} \frac{\partial E}{\partial \mu} \right] \qquad (6\times4)$$

$$Z^{-1} = \text{diagonal} \left[w_r^{-1}, w_{\dot{r}}^{-1}, w_A^{-1}, w_E^{-1} \right] \qquad (4\times4)$$

[For consistency with the notion in Equation (7), r_i could be written λ_{m_r} and the column $\partial r/\partial \mu$ is summed over the m_r values of r. The other columns of P are treated analogously.]

The computational procedure is as follows. The matrices Q(6x6), P(6x4), Z(4x4, diagonal) are kept in storage and added to as each point of data is processed. When A_t is desired, the sums are combined according to Equation (8) to form A_t^{-1}, and then A_t is computed by inversion. Additional data are processed by augmenting the appropriate sums until the next time A_t is desired, when Equation (8) is applied again. Since an initial estimate of μ, with covariance matrix A_0, is not correlated with the data processed later, it can be treated by initializing the Q sum with A_0^{-1}. [This and the corresponding procedure for Equation (1) are proved in the next paragraph.]

One further point must be considered. When the sums Q and P are incremented, the sum and the increment must be referenced to the same instant of time. This is accomplished by using the transistion matrix which is defined as

$$\varphi(t_1, t_2) = \frac{\partial \mu(t_1)}{\partial \mu(t_2)} = \frac{\partial \mu_1}{\partial \mu_2}$$

Thus, if the sums are valid at t_1 and observations λ are taken at t_2, P_2 is computed; i.e., $P(t_2) = (\partial\lambda/\partial\mu_2)^T N$, and P_1 is found as follows:

$$\begin{aligned} P_1 = P(t_1) &= \left(\frac{\partial \lambda}{\partial \mu_1}\right)^T N \\ &= \left(\frac{\partial \lambda}{\partial \mu_2} \frac{\partial \mu_2}{\partial \mu_1}\right)^T N \\ &= \left(\frac{\partial \mu_2}{\partial \mu_1}\right)^T \left(\frac{\partial \lambda}{\partial \mu_2}\right)^T N \\ &= \varphi^T(t_2, t_1) P_2 \end{aligned} \qquad (9)$$

Similarly,

$$Q_1 = \varphi^T(t_2, t_1) Q_2 \varphi(t_2, t_1) \qquad (10)$$

To update the sums from t_1 to t_2, Equation (9) and (10) are replaced by

$$P_2 = \varphi^{-1T}(t_2, t_1) P_1$$

$$Q_2 = \varphi^{-1T}(t_2, t_1) Q_1 \varphi^{-1}(t_2, t_1)$$

Since φ is a symplectic matrix, its inverse can be obtained by rearranging its terms. However, φ is obtained by numerical integration and is thus only approximately symplectic. For this reason the interval between t_1 and t_2 should not be too large in the initial portion of the trajectory, where the curvature is high, or numerical difficulties may result.

2.2 Inclusion of Initial Estimates

Consider the initial estimate to be an unbiased observation of the state vector with covariance matrix A_0. The matrices in Equation (1) for this case can be denoted by primes and partitioned as follows:

$$J' = \frac{\partial(\mu, \lambda)}{\partial \mu} = \begin{pmatrix} I \\ \cdots \\ J \end{pmatrix}$$

$$W' = \begin{pmatrix} A_0^{-1} & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & W \end{pmatrix}$$

$$N' = \frac{\partial(\mu, \lambda)}{\partial(\mu, \nu)} = \begin{pmatrix} I & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & N \end{pmatrix} \qquad (11)$$

$$\Lambda_b' = \begin{pmatrix} 0 & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & \Lambda_b \end{pmatrix}$$

$$\Lambda_r' = \begin{pmatrix} A_0 & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & \Lambda_r \end{pmatrix}$$

Note that Equation (1) could be written as

$$A_t = A^{-1}(A_r + A_b)A^{-1} \qquad (12)$$

where

$$A = J^T W J$$

$$A_b = J^T W N \Lambda_b N^T W J$$

$$A_r = J^T W \Lambda_r W J$$

(In practice the sums A, A_b, A_r are accumulated until a time t when A_t is desired.)

When Equation (11) is inserted into Equation (1), the result is

$$A_t = (A')^{-1}(A'_r + A_b)(A')^{-1} \qquad (13)$$

where

$$A' = A + A_0^{-1}$$

$$A'_r = A_r + A_0^{-1}$$

Thus the effect of an initial estimate with covariance matrix A_0 is to add A_0^{-1} to sums A and A_r and leave sum A_b unchanged. Note that if $A_0^{-1} = 0$, Equation (13) reduces to Equation (12) as expected. Also, before any data have been processed $A' = A_0^{-1}$, $A'_r = A_0^{-1}$, and Equation (13) yields $A_t = A_0$ as it should.

Similar treatment of the optimal processing case [using definitions from Equation (11)] shows that Equation (4) is replaced by

$$A_t = \left(A_0^{-1} + J^T W J\right)^{-1} \qquad (14)$$

Proceeding with Equation (14) rather than with Equation (4) it can be seen that Equation (8) must be modified by initializing Q with A_0^{-1}; i.e.,

$$A_t^{-1} = Q' - PZ^{-1}P^T \qquad (15)$$

where

$$Q' = Q + A_0^{-1}$$

2.3 Midcourse Corrections

The effect of a corrective maneuver upon the covariance matrix of the position and velocity vector of the spacecraft will be taken up next. Assume that the corrective thrust is approximately constant, that the velocity change is small compared to the vehicle velocity, and that the duration of the maneuver is small. Thus the midcourse correction can be treated as an impulse, and it can be shown that its effect is to change the covariance matrix according to

$$A_{t_+} = A_{t_-} = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & C_v \end{array}\right] \qquad (16)$$

where t_- and t_+ refer to the instant before and after the correction, 0 is the (3x3) null matrix, and C_v is the (3x3) covariance matrix of the vector of velocity increments. Assume that C_v arises from two error sources: error in the measurement of the velocity increments along the accelerometer axes, and errors in knowledge of the orientation of these axes in the inertial coordinate system. If the nominal values of the components of the incremental velocity are denoted by ξ, η, ζ, the uncertainty in the velocity increments by σ_v, and the uncertainty in the reference directions by σ_d, then

$$C_v = \sigma_v^2 I + \sigma_d^2 \begin{bmatrix} \eta^2 + \zeta^2 & \xi\eta & \xi\zeta \\ \xi\eta & \xi^2 + \zeta^2 & \eta\zeta \\ \xi\zeta & \eta\zeta & \xi^2 + \eta^2 \end{bmatrix} \qquad (17)$$

This is discussed in detail in Reference 1, pp. A-53 to A-66. The effect of Equations (16) and (17) on Equation (8) will be considered.

It would be possible to proceed as follows if the data before and after the midcourse correction were uncorrelated:

1. Compute A_{t_-} by inversion of Equation (8)
2. Compute A_{t+} using Equations (16) and (17)
3. Process data after t_+ using Equation (15) where the covariance matrix of the "initial estimate" is A_{t_+}.

Unfortunately, there is no reason why the data processed before and after the midcourse correction should be uncorrelated, and in the case under consideration, where there are bias errors in the measuring devices which are obviously unaffected by the correction, there will be inter-batch correlation. This case must be treated by altering the sums Q and P with C_v.

The correct procedure can be obtained by augmenting the state vector to include the bias sources. Then, let y be the augmented state vector where

$$y = \left(\frac{\mu}{\nu}\right) \quad \begin{array}{l} \mu \text{ is } 6\text{x}1 \\ \nu \text{ is } 4\text{x}1 \end{array}$$

The inverse covariance matrix of y is known to be (see Reference 2):

$$\text{cov}^{-1}(y) = \left(\frac{\partial\lambda}{\partial y}\right)^T \Lambda_r^{-1} \left(\frac{\partial\lambda}{\partial y}\right) + \begin{array}{l}\text{inverse covariance} \\ \text{matrix of initial estimate}\end{array}$$

where Λ_r is the covariance matrix of the observational errors. The various matrices are partitioned as follows:

$$\text{cov}^{-1}(y) = \left(\begin{array}{c|c} \left(\frac{\partial\lambda}{\partial\mu}\right)^T & \\ \hline \left(\frac{\partial\lambda}{\partial\nu}\right)^T & \end{array}\right) \Lambda_r^{-1} \left(\frac{\partial\lambda}{\partial\mu} \,\middle|\, \frac{\partial\lambda}{\partial\nu}\right) + \left(\begin{array}{c|c} A_0^{-1} & 0 \\ \hline 0 & \Lambda_b^{-1} \end{array}\right)$$

$$= \left(\begin{array}{c|c} J^T & \\ \hline N^T & \end{array}\right) \Lambda_r^{-1} (J \mid N) + \left(\begin{array}{c|c} A_0^{-1} & 0 \\ \hline 0 & \Lambda_b^{-1} \end{array}\right)$$

$$= \left(\begin{array}{c|c} J^T\Lambda_r^{-1}J + A_0^{-1} & J^T\Lambda_r^{-1}N \\ \hline (J^T\Lambda_r^{-1}N)^T & N^T\Lambda_r^{-1}N + \Lambda_b^{-1} \end{array}\right)$$

$$= \left(\begin{array}{c|c} Q' & P \\ \hline P^T & Z \end{array}\right)$$

The point of all this manipulation should become clear if one notes $Q' \neq \text{cov}^{-1}(\mu)$. First, the above result will be used as an alternate derivation of Equation (1). Using Schur's identity[3] to find cov(y) in a partitioned form, one has $\text{cov}(\mu) = \Delta^{-1}$ where

$$\Delta = J^T\Lambda_r^{-1}J + A_0^{-1} - \left(J^T\Lambda_r^{-1}N\right)\left(N^T\Lambda_r^{-1}N + \Lambda_b^{-1}\right)^{-1}\left(J^T\Lambda_r^{-1}N\right)^T = Q' - PZ^{-1}P^T$$

This is equivalent to Equation (5) except that A_0^{-1} is explicitly included, as in Equation (15). Also the various blocks of $\text{cov}^{-1}(y)$ correspond to the sums in Equation (15).

To proceed further, $\text{cov}^{-1}(y)$ must be partitioned as follows:

$$\text{cov}^{-1}(y) = \left(\begin{array}{c|c|c} R_{33} & S_{33} & T_{34} \\ \hline U_{33} & V_{33} & W_{34} \\ \hline X_{43} & Y_{43} & Z_{44} \end{array}\right) \qquad (18)$$

where the number of rows and columns in each block are indicated by subscripts. The previously defined sums will be given in terms of these blocks by

$$Q'(6X6) = \left(\begin{array}{c|c} R & S \\ \hline U & V \end{array}\right) \qquad (19)$$

$$P(6X4) = \left(\begin{array}{c} T \\ \hline W \end{array}\right)$$

Also, note that $U = S^T$, $X = T^T$ and $Y = W^T$.

If cov (y) is denoted before and after the midcourse correction by $A_\pm$,

$$A_+ = A_- + \left(\begin{array}{c|c|c} O_{33} & O_{33} & O_{34} \\ \hline O_{33} & C_v & O_{34} \\ \hline O_{43} & O_{43} & O_{44} \end{array}\right) \qquad (20)$$

where $A_\pm$ in partitioned form is given in equation (18). [Compare Equation (20) for the augmented covariance matrix with Equation (16) for the unaugmented matrix.] Note that Equation (20) can be written in the form

$$A_+ = A_- + BC_vB^T \qquad (21)$$

where $B^T = (O_{33} \mid I_{33} \mid O_{34})$. Use of the matrix inversion lemma on Equation (21) yields

$$A_+^{-1} = A_-^{-1} - A_-^{-1}B\left(C_v^{-1}+B^TA_-^{-1}B\right)^{-1}\left(A_-^{-1}B\right)^T \qquad (22)$$

By performing the manipulations in Equation (22) and defining

$$M \equiv \left(C_v^{-1}+V\right)^{-1}$$

it is found that

$$\begin{aligned} R_+ &= R_- - S_-MS_-^T \\ S_+ &= S_- - S_-MV_-^T \\ V_+ &= V_- - V_-MV_-^T \end{aligned} \qquad (23)$$

$$\begin{aligned} T_+ &= T_- - S_-MW_- \\ W_+ &= W_- - V_-MW_- \end{aligned} \qquad (24)$$

$$Z_+ = Z_- - W_-^TMW_- \qquad (25)$$

Thus Equations (23) to (25) can be used to replace Q'_- by Q'_+, P_- by P_+ [see Equation (19)], and Z_- by Z_+, and to compute A_{t_+}. Explicitly,

$$A_{t_\pm} = Q'_\pm - P_\pm Z_\pm^{-1}P_\pm^T \qquad (26)$$

gives the inverse of the covariance matrix of μ before and after the midcourse correction. Equation (26) can no longer be used for the case of zero bias errors ($\gamma_i = 0$) as might be expected from its derivation, because $\text{cov}^{-1}(y)$ of Equation (18) becomes singular with $\Lambda_b \to 0$.

2.4 Numerical Results

A nominal trajectory was chosen from the large number available in Reference 4. It was a low-energy (vis viva integral of 8.049 km^2/s^2) trajectory with nominal flight time of 198 days. A heliocentric ecliptic map of the trajectory is shown in Figure 134.

It is assumed that the data processing would be handled in three phases: near-Earth, interplanetary, and terminal. The following discussion is concerned mainly with the interplanetary phase. The near-Earth phase (the first 10 days after injection) was considered only to obtain reasonable initial estimates of the covariance matrix for the interplanetary phase. More discussion of near-Earth tracking will follow in Section 3.

The expressions σ_p and σ_v will be used as measures of the accuracy of the orbit determination:

$$\sigma_p = \left|a_{11}^2 + a_{22}^2 + a_{33}^2\right|^{1/2}$$

$$\sigma_v = \left|a_{44}^2 + a_{55}^2 + a_{66}^2\right|^{1/2}$$

and the a_{ij} are the diagonal elements of the covariance matrix. The expressions σ_p (in thousands of feet) and σ_v (in hundredths of feet per second) are shown for the first 10 days

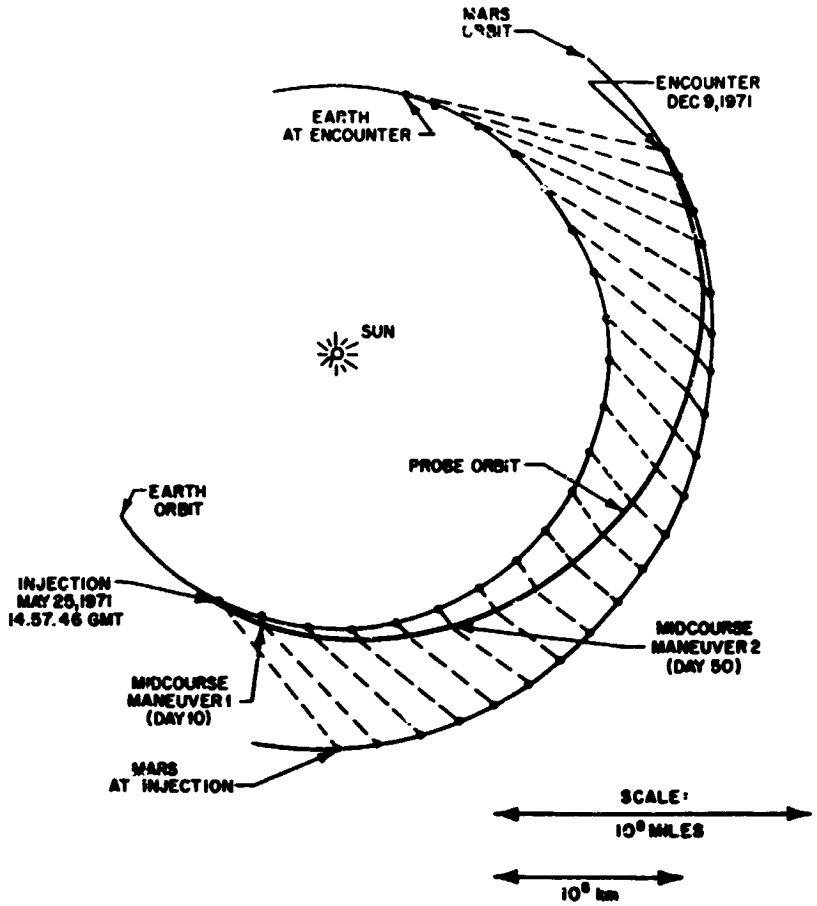


Figure 134. Trajectory to Mars - positions at 10-day intervals

in Figure 135. (Alternate scales for σ_p in kilometers and σ_v in mm/s are given on the right of all the figures.) The data were processed optimally, i.e., according to Equation (8) using the tracking parameters of Table 59. The covariance matrix at Day 10 is used as the initial estimate in the various interplanetary runs considered next.

During the near-Earth phase, only a single tracker on the Earth's equator is assumed to exist. Similarly, the measurements during the interplanetary phase are assumed to come from a single synchronous satellite.† In both cases, the location of the station is assumed to be known exactly. The effect of station location errors, which are always present in practice, is to make the uncertainties in position and velocity somewhat optimistic. There is no reason to expect that the comparative accuracies discussed below are significantly affected by station location errors. They are dependent, however, on the error characteristics assumed for the different tracking data. To arrive at a justifiable set of values for these parameters (range, range rate, and angular precision), several possibilities are examined.

Table 59

NEAR-EARTH TRACKING PARAMETERS

σ_r (feet)*	1000 (0.3)
$\sigma_{\dot{r}}$ (feet/second)*	0.1 (30)
σ_A (radians)	0.2×10^{-3}
σ_E (radians)	0.2×10^{-3}
γ_r (feet)*	1000 (0.3)
$\gamma_{\dot{r}}$ (feet/second)*	0.1 (30)
γ_A (radians)	0.2×10^{-3}
γ_E (radians)	0.2×10^{-3}
Data rate	One observation per 15 minutes

*Approximate equivalents in km and mm/s are in parentheses. As always, σ stands for the standard deviation of a random error and γ for that of a bias error. The subscripts A and E stand for measurements of azimuth and elevation angle, respectively.

Table 60 lists a range of characteristics for the measurement data serving as inputs to the tracking calculations for the interplanetary phase. In range and range rate, they vary from high-quality data, representing the best possible in the forseeable future, to inferior ones (alternatives I-II), where the degraded inputs serve as a simplified allowance for geophysical bias errors,‡ whose rigorous treatment would require a more elaborate filtering algorithm.[8] Three levels of optical precision are also introduced, varying from standard deviations of 2 s, representing exoatmospheric observations, to 15 s, for endoatmospheric readings; i.e., the former apply for tracking relays on board Earth satellites or at the libration points while the latter are representative of unpredictable errors at ground observatories. In every case the assumed value of the measurement bias (i.e., the γ, to be distinguished from geophysical bias) is set equal to the assumed value of the random measurement error (i.e., the σ). Of course, their exact relation depends on the specific apparatus under consideration. In the absence of any particular instrumentation to which to tailor this study, the assumption $\sigma = \gamma$ seems as good as any.

Table 60

VARIATION OF DEEP SPACE TRACKING PARAMETERS

I	σ_r = 328 ft (100 m)
	$\sigma_{\dot{r}}$ = 0.003 fps (1 mm/s)
II	σ_r = 3280 ft (1 km)
	$\sigma_{\dot{r}}$ = 0.006 fps (2 mm/s)
III	σ_r = 328,000 ft (100 km)
	$\sigma_{\dot{r}}$ = 0.01 fps (3 mm/s)
a	$\sigma_A = \sigma_E = 10^{-5}$ radians (2s)
b	$\sigma_A = \sigma_E = 1.75 \times 10^{-5}$ radians (3.5s)
c	$\sigma_A = \sigma_E = 2.5 \times 10^{-5}$ radians (5s)
d	$\sigma_A = \sigma_E = 7 \times 10^{-5}$ radians (15s)

Bias errors: for all cases $\gamma_r = \sigma_r$ etc.

Data rate: one observation per 12 hours

One source of geophysical bias is the errors in station location, including survey errors, wandering of the poles, and non-uniformities of the Earth's rate of rotation.[5,7] The main concern, however, is the uncertainty in the speed of light, c, of 2×10^{-6}. This corresponds to an uncertainty in the astronomical unit, A, of several hundred kilometers even though the light time, τ_A (number of light seconds in one astronomical unit), is known much better – with a relative uncertainty of 0.01×10^{-6}. Indeed, one might take the light time as the fundamental unit of length, and thus make other (terrestrial) units extraneous.[5] However, as long as one does not do so it seems unrealistic to take one's precision of measurement much greater than the uncertainty

†These single-tracker situations do not take advantage of the operating modes discussed in Section 3, but this simplification does not detract from the comparison of different kinds of tracking data for the interplanetary phase.

‡Thus, for example, $\sigma_{\dot{r}}$ = 3 mm/s is considered typical of current DSIF tracking.[5]

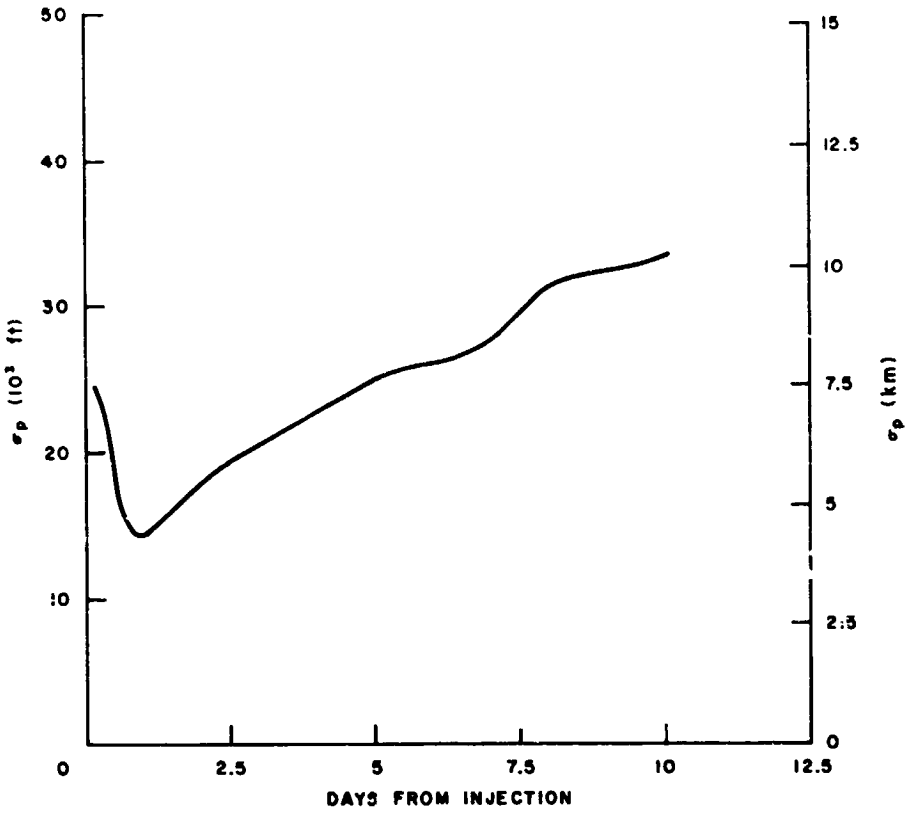


Figure 135a. Near-Earth phase

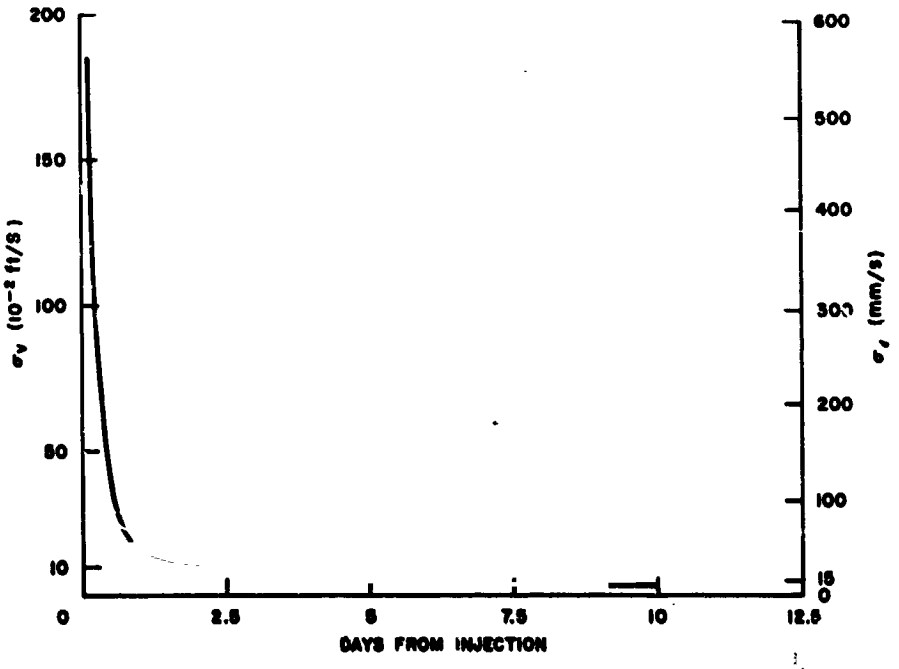


Figure 135b. Near-Earth phase

of the unit it is measured in. The uncertainty in c affects both the range and range rate data. The corresponding uncertainty in A is taken from Reference 5, and is the value obtained after data from recent interplanetary missions were processed. Thus it may not decrease drastically in the near future. However, sets I and II are included to show the effects of eventual improvements in the astronomical unit A.*

Figure 136 considers the effects of range and range rate data of different accuracy. The results were obtained using the parameters listed as alternatives I, II, and III in Table 60. On Fig. 136a an initial error buildup in σ_p is noted for the less accurate tracking data represented by II and III. As tracking data accumulate, this error growth is eventually overcome and the usual asymptotic behavior prevails. With alternative I, representing tracking accuracies expected some time in the future, this transient never occurs. In all cases the σ_v plots of Figure 136b show a purely monotonic trend.

Figures 137 to 139 show what happens when angular data are added to the r, $\dot{r}$ observations. It is recalled that alternative (a) of Table 60 is typical of high-grade optical tracking devices, relying on stellar references. Alternative (d) is representative of the accuracy to be expected from Earth-based optical trackers, which are subject to atmospheric perturbations. In the latter case the error propagation is not significantly affected by the optical data, but the high-grade optical measurements from alternative (a) prevent an initial error buildup if used with r, $\dot{r}$ data representative of the current state of the art. Thus it is the ratios of the σ's (and γ's) of different types of measurement that are significant. In view of the uncertainties in terrestrial and astrophysical constants, which affect the r, $\dot{r}$ data as mentioned earlier, the combination III(a) of tracking accuracies is felt to be a realistic one under the present circumstances, and one that shows some benefit from optical data.

One parameter, the data rate, should be considered somewhat further. A rate of 1 point every 12 hours has been assumed for the interplanetary phase of the flight. This appears to be reasonable for data-processing purposes. However, it is interesting to consider the effect of other rates. Processing more data can have two favorable effects upon the variances. The first is simply the standard phenomenon of reducing the variance of some variable by taking more measurements; this is sometimes called the "square-root effect." The second improvement is due to the fact that the geometry changes with time and that the additional measurements are qualitatively different. One would expect this geometric effect to become negligible at some small enough data interval, and indeed this can be demonstrated.

Inspection of Equation (7), for example, will show that if there is no geometric effect, i.e., if the derivatives $\partial\lambda/\partial\mu$ do not vary with time, then if one takes n times as many measurements A_t^{-1} will increase by a factor of n. Thus A_t will decrease by n and σ_p and σ_v will decrease by a factor of $\sqrt{n}$.† The effect of actually increasing the data rate by factors of 4 from 1 every 48 hours to 1 every 45 minutes is shown in Table 61. It can be seen that the geometric effect becomes negligible somewhere between 1 every 12 hours and 1 every 3 hours, and that data rates higher than this benefit only from the square-root effect. This will eventually drive the estimation error variances to zero as the data rate approaches infinity for the error model we have considered. For other types of bias errors, where the present estimator may no longer be optimal, or for least-square estimation as discussed below, this may no longer be true. For a complete study of these questions rather detailed instrument characteristics and peculiarities of the measurement process must be taken into account, some of which are discussed in the specialist literature.‡ Whatever a useful upper bound for the data rates in r and $\dot{r}$ turns out to be, it is usually true that increased rates for this kind of data produce the same beneficial effects as a result from the addition of angle measurements.

The effect of optimal vs. weighted least-squares data processing is demonstrated in Figures 140 (for r, $\dot{r}$ and angles) and 141 (for r, $\dot{r}$ alone). The effect is seen to be large, with the position uncertainty at encounter being greater by a factor of 8 with the weighted least-squares method.

Of course, the results with the weighted least-squares method depend upon the choice of weights, i.e., of the diagonal elements of W in Equation (1). This choice appears to be somewhat of an art rather than a science. One rule of thumb, used by the developers of Reference 1, was to take[9]

$$w_i^2 = \sigma_i^2 + m_i\gamma_i^2$$

where m_i is the number of measurements. If $\sigma_i = \gamma_i$ and $m_i \approx 10^3$, this results in

$$w_i \approx (m_i)^{1/2}\sigma_i$$

*Since $A = c\tau_A$, the errors in these quantities must be consistent according to that equation. In Reference 5, for example, c is held fixed, and A and τ_A are solved for as part of the post-flight data analysis. Clearly, the consistent improvement of the systems of units employed in astronomy and physics, with the help of data from deep space missions, is a highly complex matter. It involves the fitting of statistical models of vehicle dynamics, measurement, and computing procedures to a great variety of observational data. A detailed discussion in the present context will not be possible.

†When the data rate is increased by n, one must also replace A_0^{-1} by nA_0^{-1} and γ_i by $(1/\sqrt{n})\gamma_i$, to get a factor of $\sqrt{n}$ exactly in the sigmas.

‡For example, in Reference 6 note that JPL's tracking rates for the lunar orbiter missions are 1/min in Doppler and 1/3 min in range. Presumably these are optimal rates chosen as a compromise between various system considerations for this type of mission.

Table 61

EFFECT OF DATA RATE

Data Rate	1/48 hours	1/12 hours	1/3 hours	1/45 minutes
σ_p (feet)* at Day 50	202,089 (62)	25,883 (7.9)	7,368 (2.2)	3,687 (1.1)
σ_v (fps)* at Day 50	0.03935 (12)	0.00532 (1.6)	0.00154 (0.5)	0.00077 (0.25)
Ratio of σ to σ at 1/4 Data Interval	→ 7.5	→ 3.5	→ 2.0	

*Approximate equivalents in km and mm/s in parentheses.

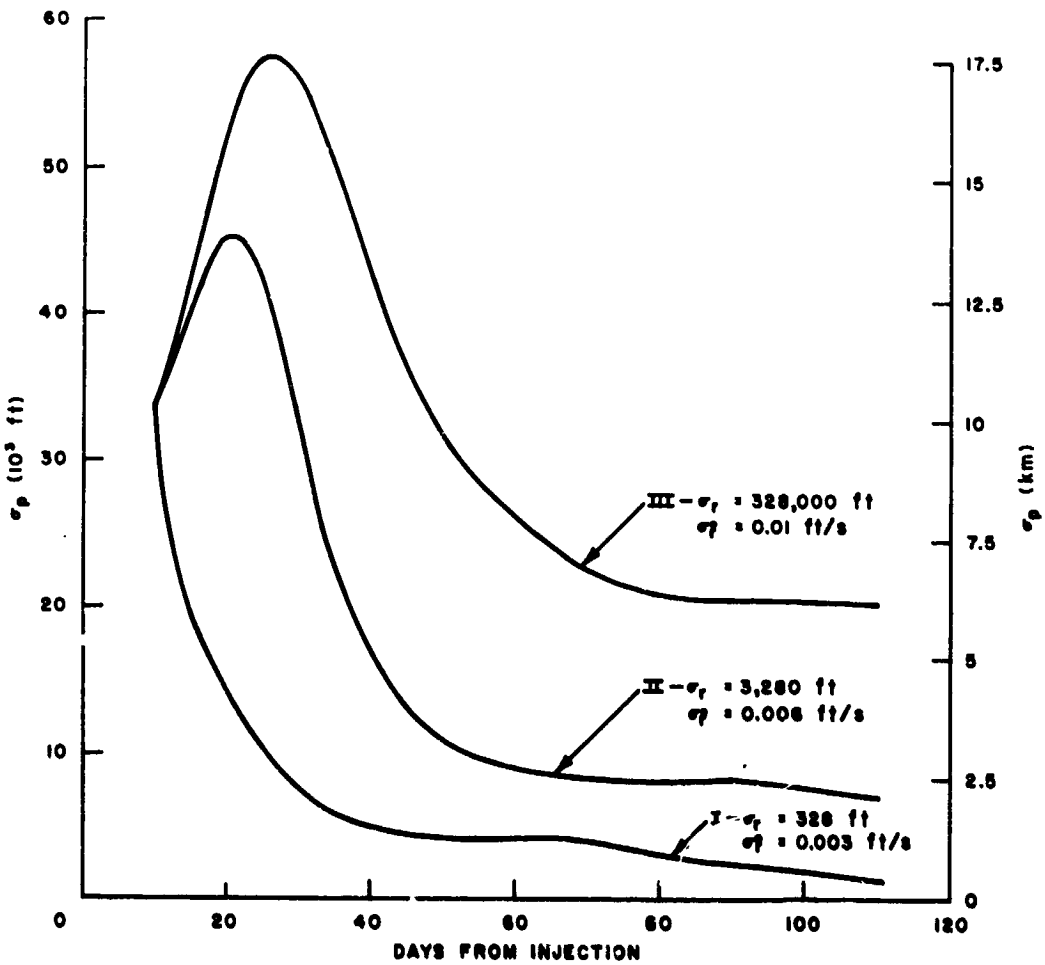


Figure 136a. Effect of range and doppler accuracy

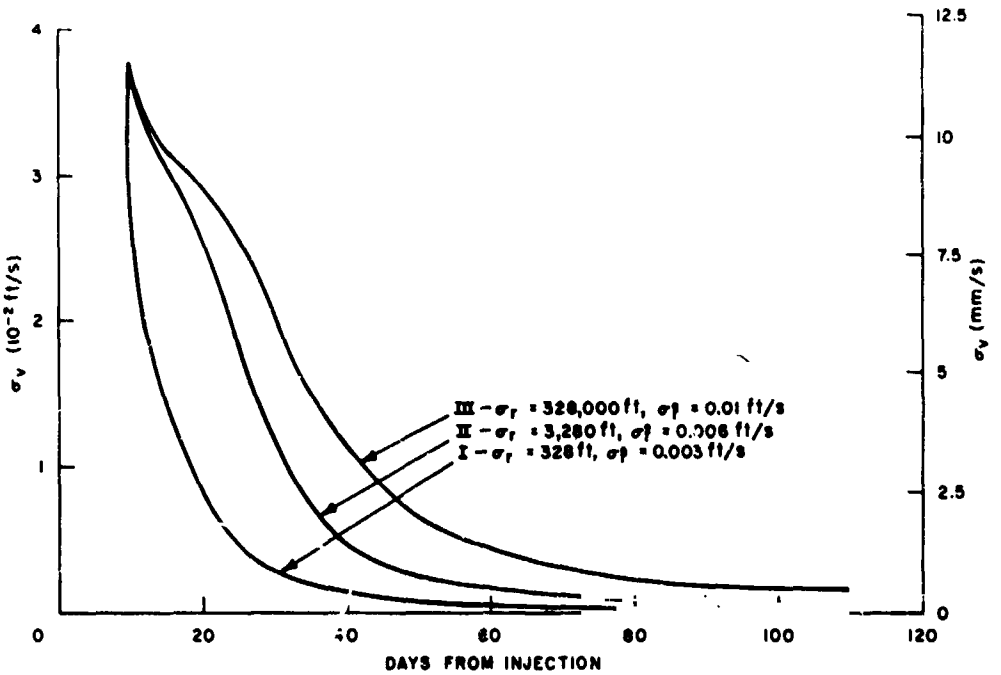


Figure 136b. Effect of range and doppler accuracy

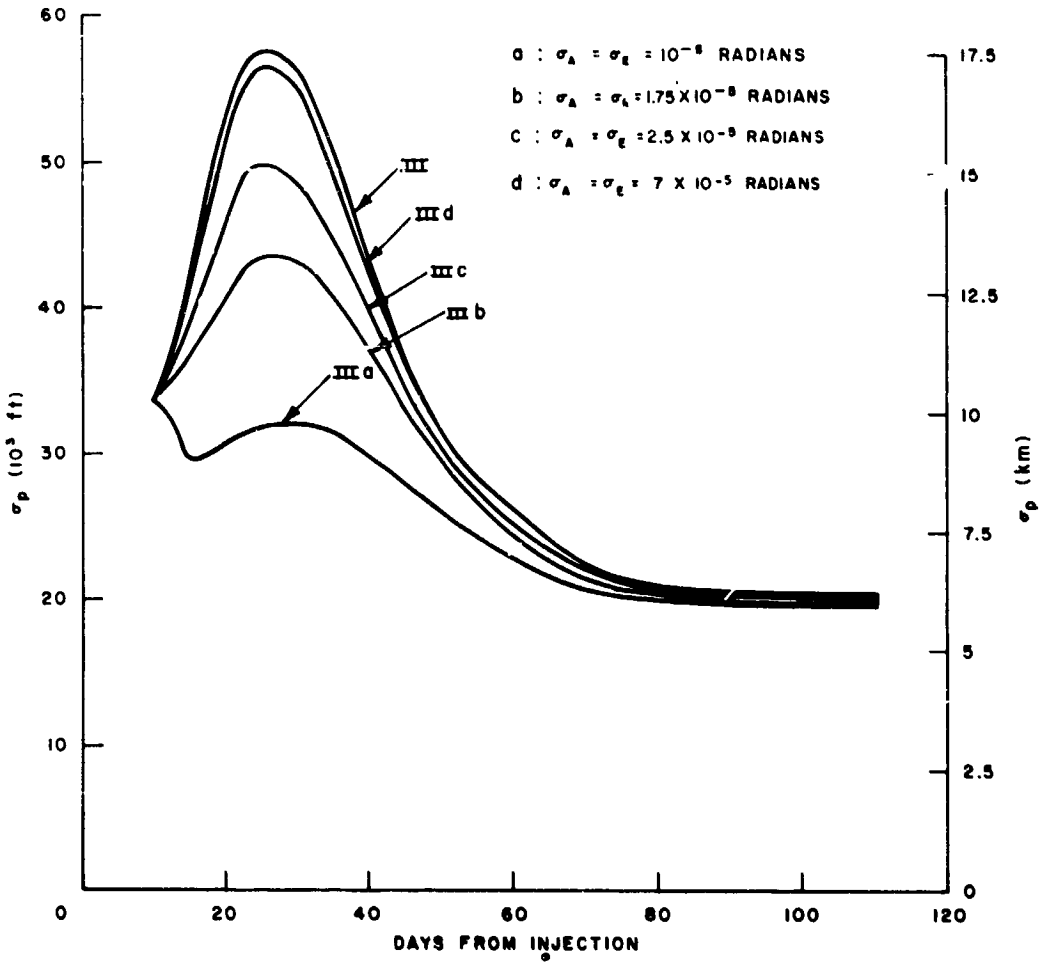


Figure 137a. Effect of angles (III: σ_r = 328,000 ft, $\sigma_{\dot{r}}$ = 0.01 ft/s)

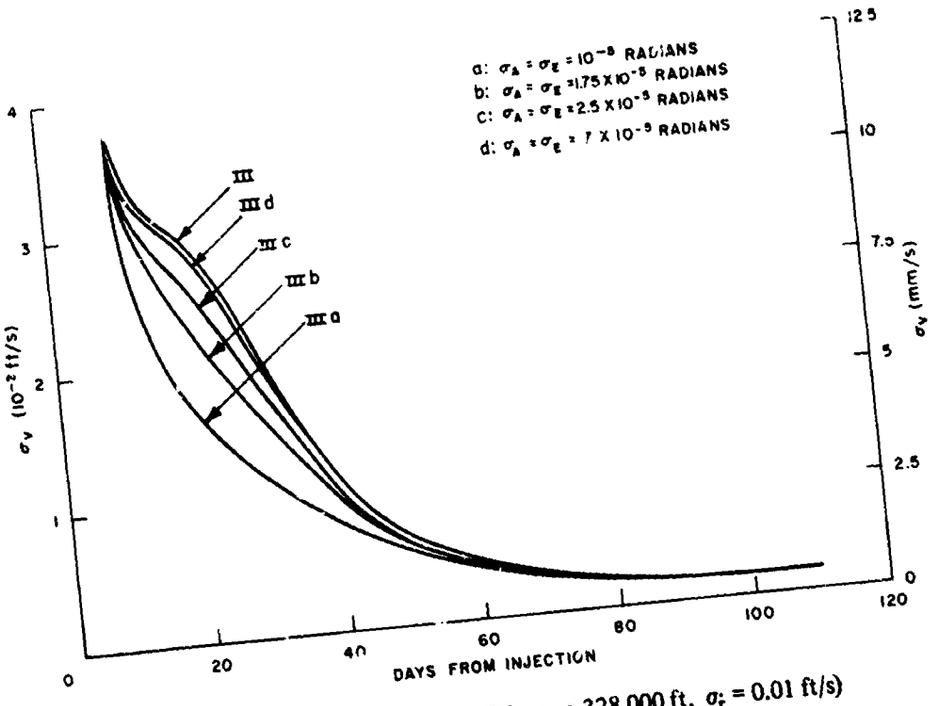


Figure 137b. Effect of angles (III: $\sigma_r = 328{,}000$ ft, $\sigma_{\dot{r}} = 0.01$ ft/s)

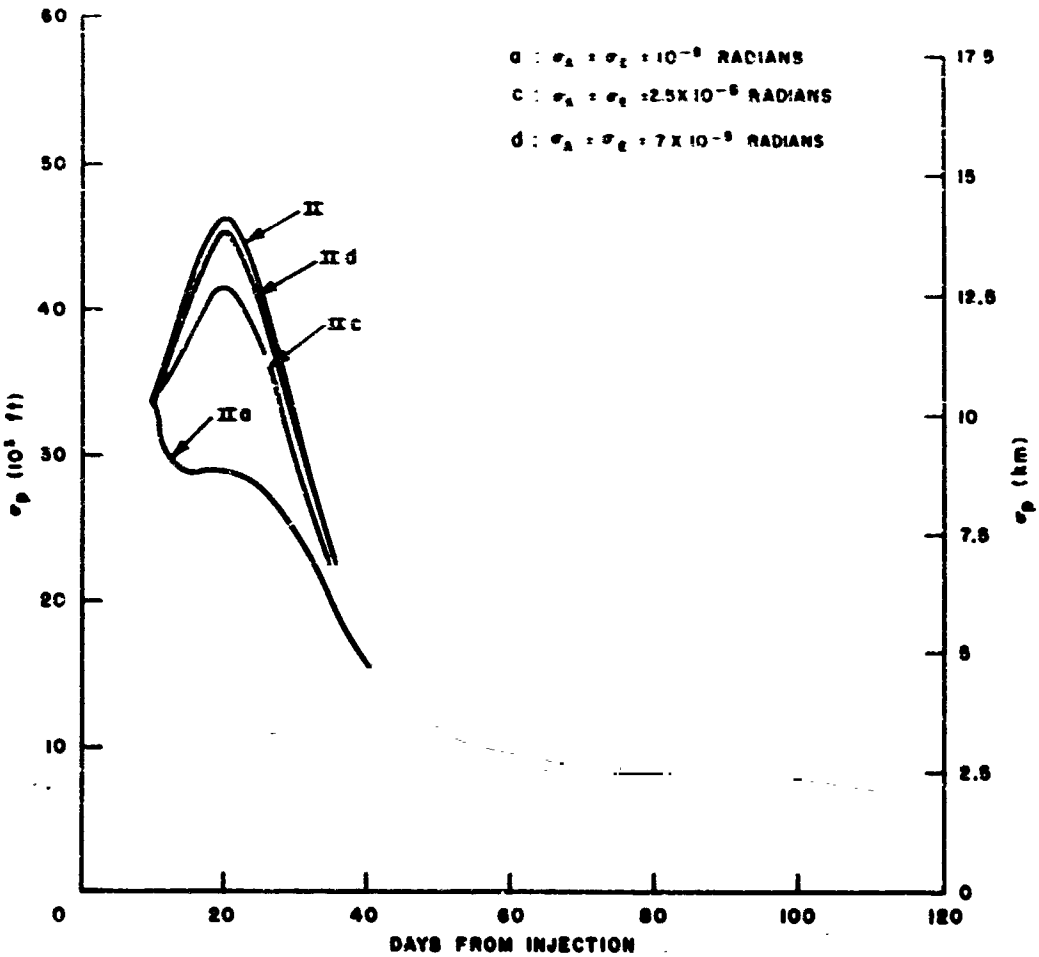


Figure 138a. Effect of angles (II: $\sigma_r \approx 3{,}280$ ft, $\sigma_{\dot{r}} \approx 0.006$ ft/s)

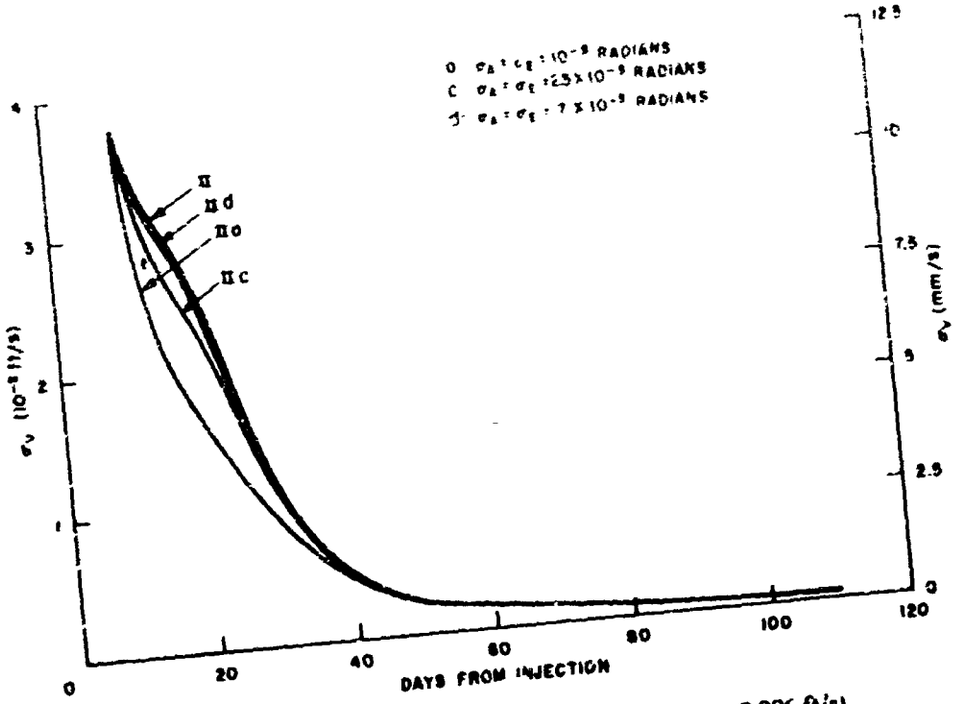


Figure 138b. Effect of angles (II: $\sigma_r = 3{,}280$ ft, $\sigma_{\dot{r}} = 0.006$ ft/s)

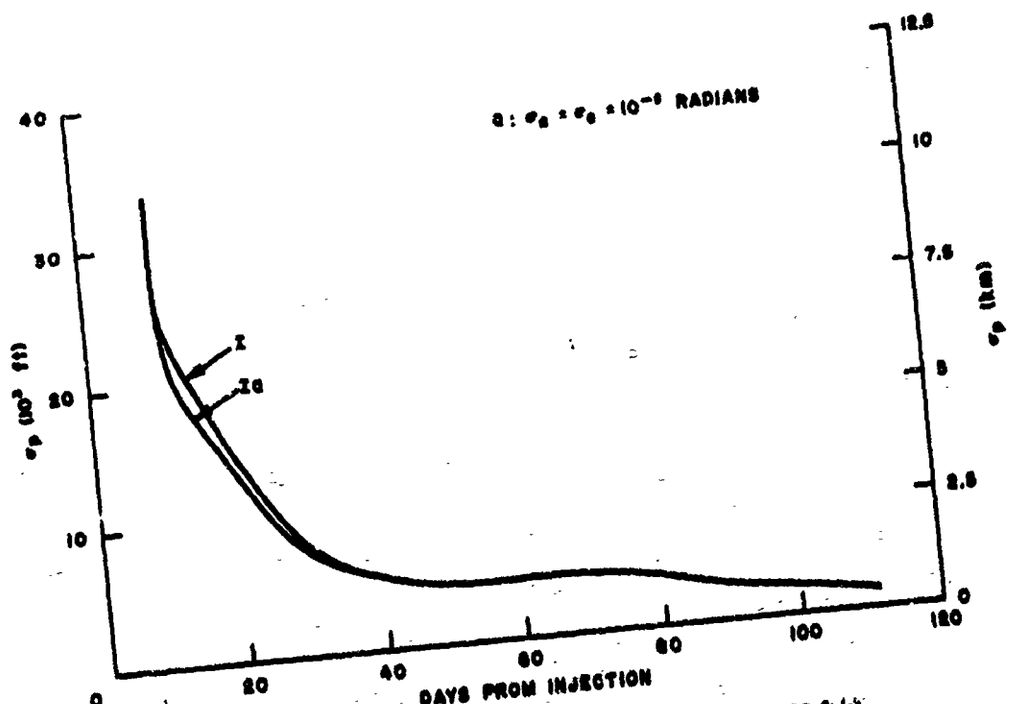


Figure 139a. Effect of angles (I: $\sigma_r = 328$ ft, $\sigma_{\dot{r}} = 0.003$ ft/s)

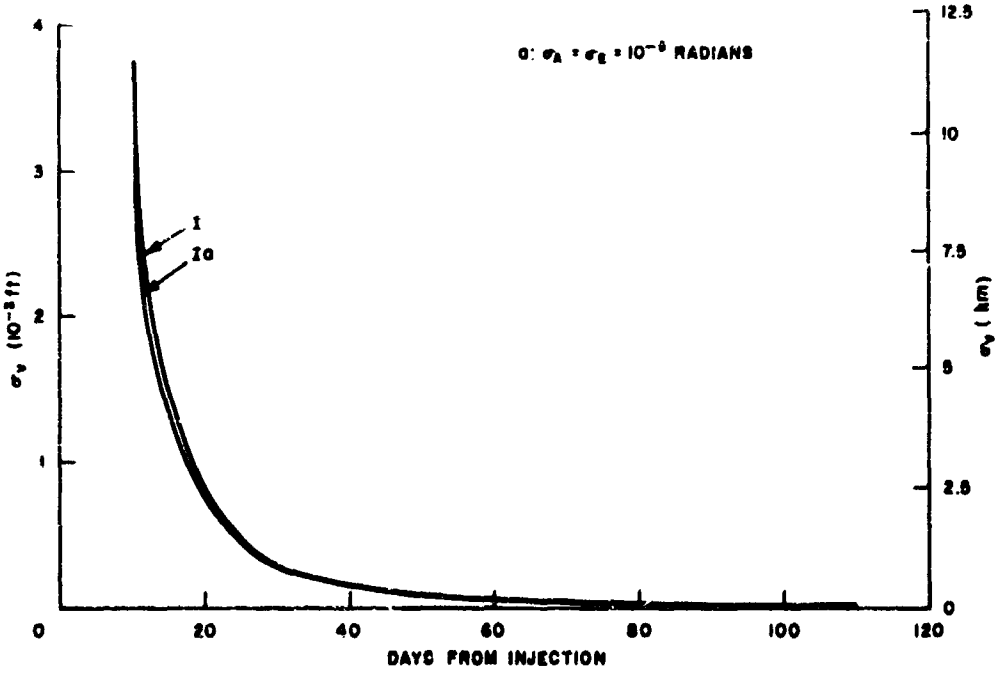


Figure 139b. Effect of angles (I: $\sigma_r = 328$ ft, $\sigma_{\dot{r}} = 0.003$ ft/s)

Possibly other choices of W will give better results, but of course none can give results better than the optimal filter.

The least-squares results with and without angles (see Figures 140 and 141) are not significantly different, unlike the optimal processing results. This is possibly a result of the particular weighting used. The terminal rise in the σ_v plots, which is particularly pronounced for the least-square processor, reflects the perturbations from Mars as the vehicle enters the planet's near field.

The position and velocity uncertainties after midcourse corrections with and without accurate angular data are considered next (see Figure 142). Midcourse corrections with components of ξ, η, and ζ in Equation (17) of 30 fps (9 m/s) are applied at 10 and 50 days. (Actually, 50 days is late in practice for the second midcourse correction. It is used here to emphasize the influence of the time of midcourse correction upon the effect of inclusion of angular data.) It is assumed that the uncertainty in the velocity increments is 1 part in 1000, or 0.03 ft/s (9mm/s), and the uncertainty in the reference directions 0.2×10^{-3} radians or about 0.01 degrees. These result in an increase in velocity uncertainty of 0.0030 ft/s (1 mm/s) and 0.0043 ft/s (1.3 mm/s) at 10 and 50 days, respectively.

The angular data reduce the sigmas significantly for the correction at Day 10, but much less for the one at Day 50. In fact, it appears likely that the difference in σ_p between the two curves after Day 50 (Figure 142a) is due in large measure to the fact that the effect of the first midcourse correction on σ_p is still non-negligible when the second correction is applied. Note again the terminal growth of σ_v due to Martian gravity.

2.5 Conclusions

The presence of highly accurate angular data, such as would be obtained from optical instrumentation outside the atmosphere, reduces position and velocity uncertainties significantly during the early portion of an interplanetary flight in the presence of r, $\dot{r}$ data representative of current tracking techniques. Thus, for example, the addition of optics with $\sigma = 2$ s to radar data with $\sigma_r = 1$ km and $\sigma_{\dot{r}} = 2$ mm/s reduces the position error σ_p from 13 to 8 km and the velocity error σ_v from 7.5 to 5 mm/s after the first 10 days of tracking. Similar results occur after a midcourse correction. The effect is much less in the later portion of the flight. Use of optimal data processing (as opposed to weighted least squares) is shown to very significantly reduce position and velocity uncertainties for the entire duration of the flight. The extremely small errors which are theoretically demonstrated suggest that relatively minor geophysical uncertainties, such as the earth's rotation, and polar motion, need to be investigated to actually obtain these accuracies (References 5 to 7). This will require elaborate mathematical models for various types of bias errors, the proper treatment of correlations in random error components, and an allowance for the fact that σ_r and $\sigma_{\dot{r}}$ are actually functions of range. Such considerations may go hand in hand with astrometric* measurements of the space probe from near-Earth stations as an ultimate refinement of optical techniques.

3. NAVIGATION NEAR THE EARTH

In this section an error propagation study is performed with methods similar to those described in Section 2, to reveal the salient properties of interplanetary navigation schemes near the Earth. Range, range rate, and angular data from multiple tracking stations are considered. In particular, it is assumed that Earth satellites and stations at the Earth-Moon libration points perform the tracking function. A two-dimensional model of this situation is developed to compare the merits of optical angle data and trilateration range measurements.

3.1 Equations of Motion

The space probe will be considered as a point mass moving in field-free space. Its trajectory is approximated as a straight line and a planar geometry is assumed. This greatly simplified model is still an adequate representation of the salient features of near-Earth tracking. The geometry of this tracking situation is shown in Figure 143.

The dynamical equations are given in terms of the vehicle position coordinates $\bar{x}$, $\bar{y}$ as

$$\ddot{\bar{x}} = f_x ; \ddot{\bar{y}} = f_y \tag{27}$$

The forces (or applied accelerations) f_x, f_y are to represent possible corrective thrust terms.

The solution to Equation (27) with $f_x = f_y = 0$ is

$$\begin{aligned} \bar{x} &= a_1 + a_3 (t-t_0) \\ \bar{y} &= a_2 + a_4 (t-t_0) \\ \dot{\bar{x}} &= a_3 \\ \dot{\bar{y}} &= a_4 \end{aligned} \tag{28}$$

*The astrometric method consists of measuring the distances between the image of a target and neighboring identifiable stars on some form of semi-permanent record (such as a photographic emulsion or an image orthicon tube). These measurements tend to be more precise than angle read-outs from the pointing optics in real time.

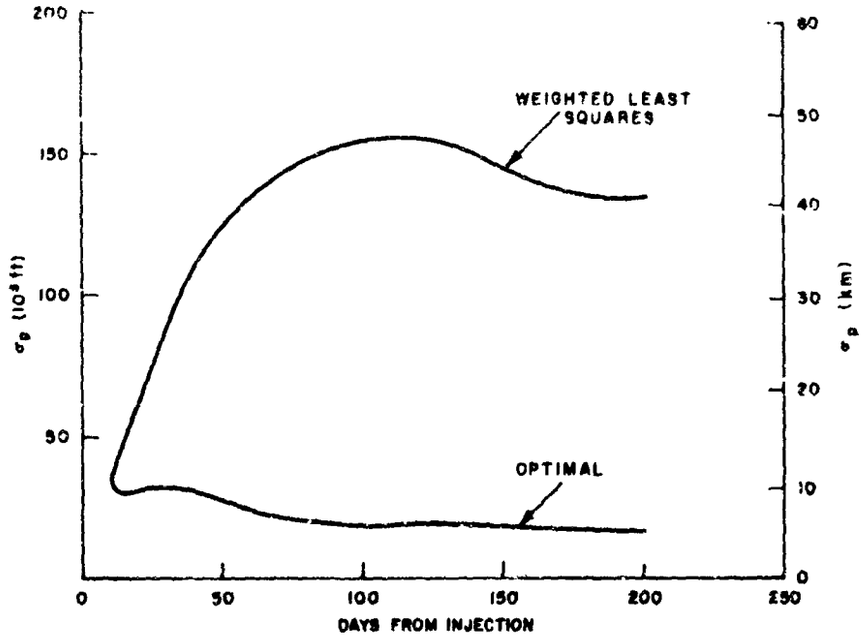


Figure 140a. Effect of type of data processing with r, $\dot{r}$, A, E data

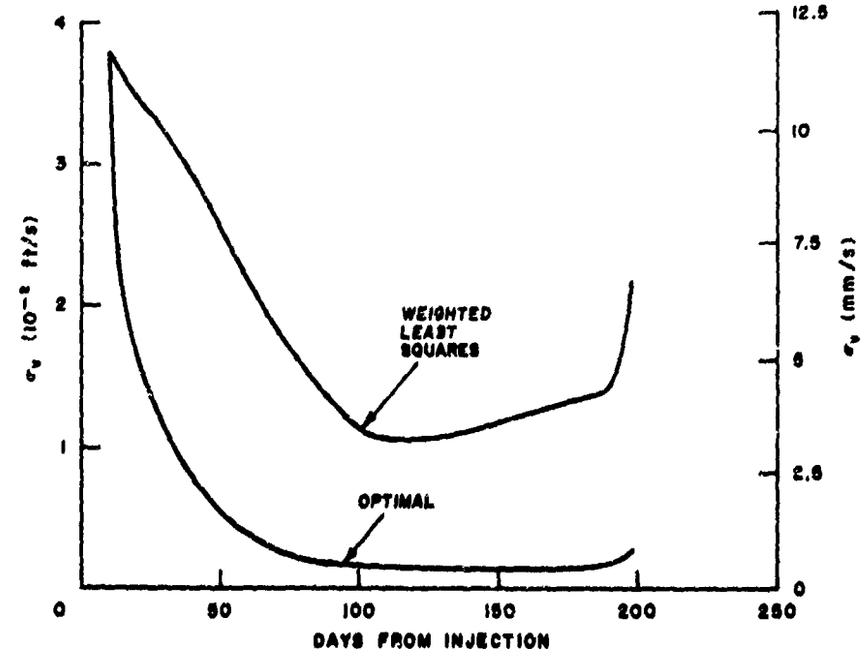


Figure 140b. Effect of type of data processing with r, $\dot{r}$, A, E data

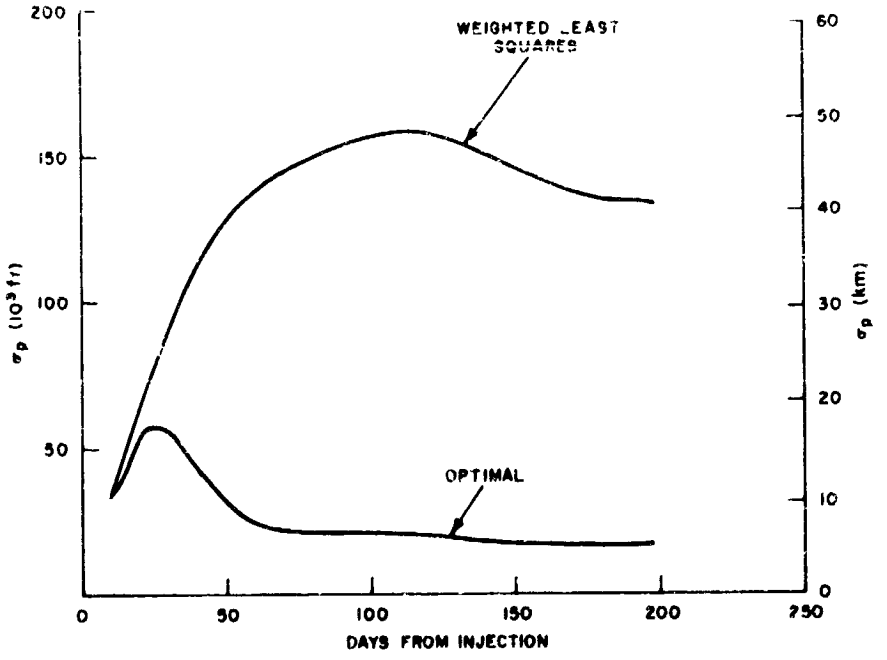


Figure 141a. Effect of type of data processing with r, $\dot{r}$ data

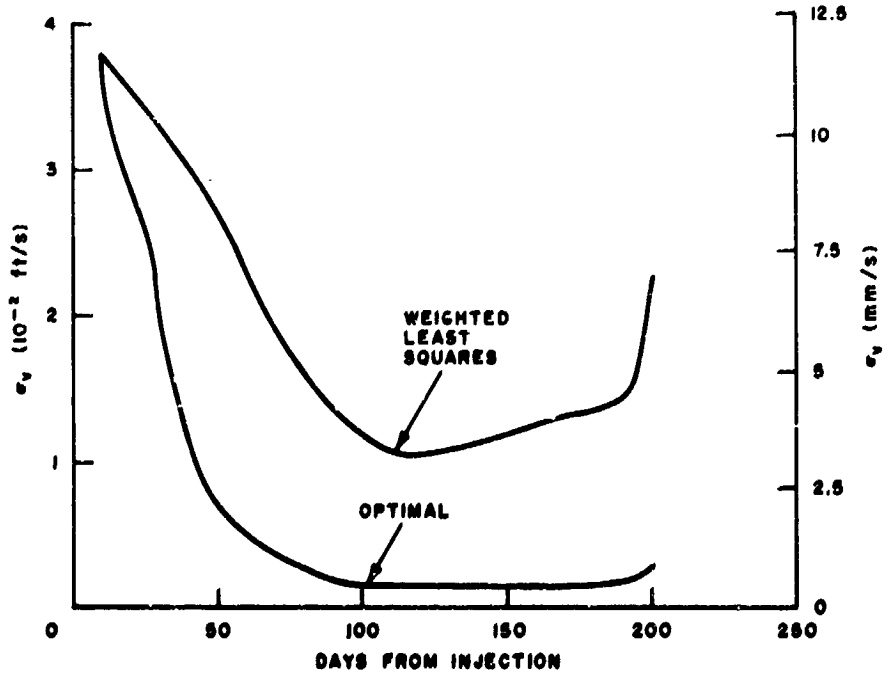


Figure 141b. Effect of type of data processing with r, $\dot{r}$ data

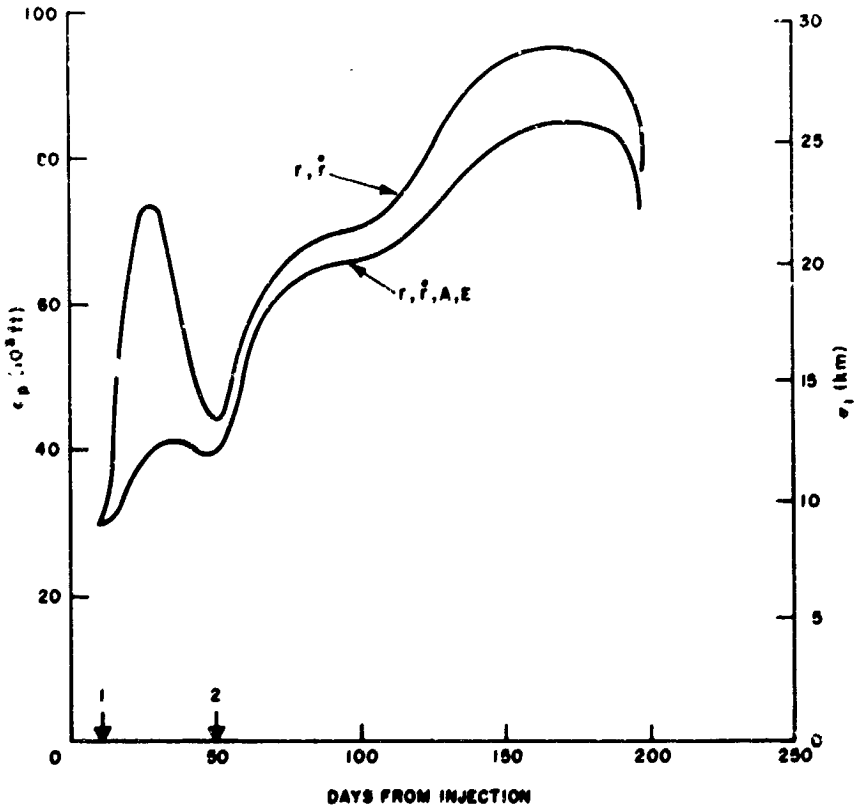


Figure 142a. Effect of mid-course corrections at 10 and 50 days

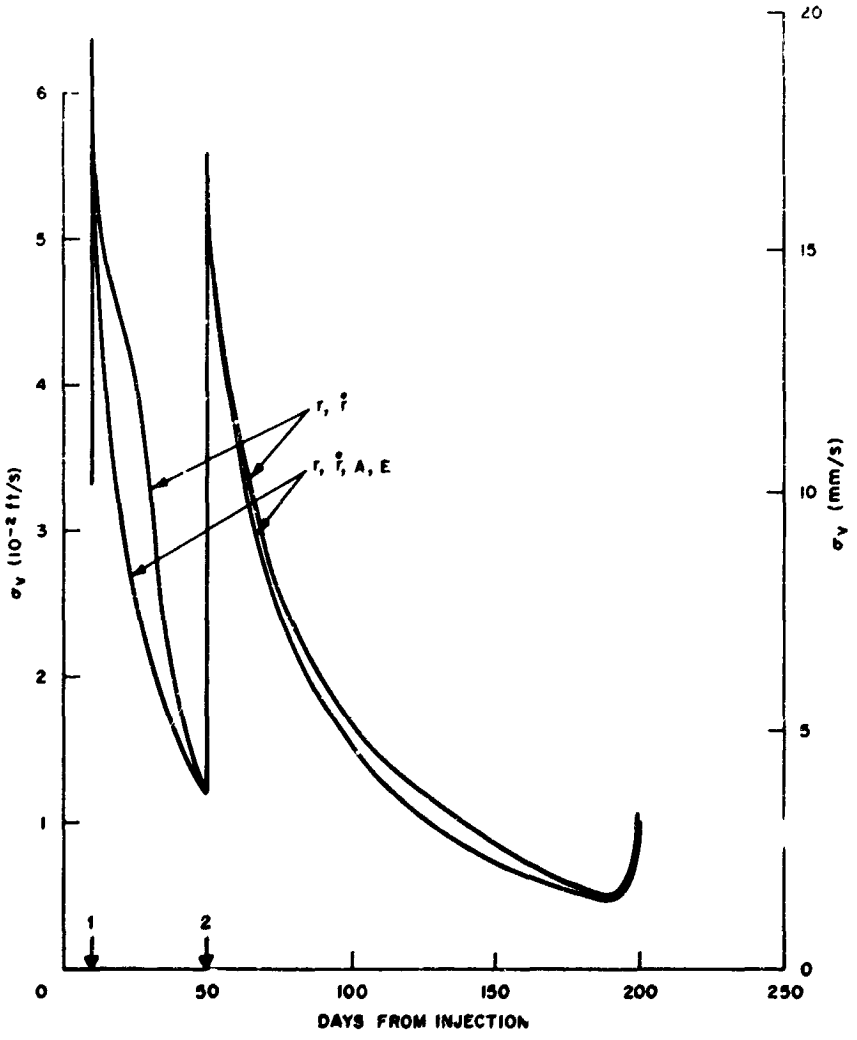


Figure 142b. Effect of mid-course corrections at 10 and 50 days

The parameters a_1, a_2, a_3, a_4 are the "orbital elements" of this system. The time variation of the elements follows the equations

$$\begin{aligned} \dot{a}_1 &= -f_x\,(t-t_o) \\ \dot{a}_2 &= -f_y\,(t-t_o) \\ \dot{a}_3 &= f_x \\ \dot{a}_4 &= f_y \end{aligned} \tag{29}$$

3.2 Observations

Observations are considered with respect to a station, designated by the subscript i, at position $\bar{x}_i$, $\bar{y}_i$. The observed range, range rate, and angle are

$$\begin{aligned} r_i &= \left(x_i^2 + y_i^2\right)^{1/2} \\ \dot{r}_i &= \left(x_i\dot{x}_i + y_i\dot{y}_i\right)\left(x_i^2 + y_i^2\right)^{-1/2} \\ \theta_i &= \tan^{-1}\left(y_i/x_i\right) \end{aligned} \tag{30}$$

where

$$\begin{aligned} x_i &= \bar{x} - \bar{x}_i \\ y_i &= \bar{y} - \bar{y}_i \\ \dot{x}_i &= \dot{\bar{x}} - \dot{\bar{x}}_i \\ \dot{y}_i &= \dot{\bar{y}} - \dot{\bar{y}}_i \end{aligned}$$

A known nominal trajectory $\bar{x}$, $\bar{y}$ is considered along with deviations from the nominal, δx, δy. The corresponding deviations in the observations made from the i^{th} station are z_i, where

$$z_i = \begin{bmatrix} \delta r_i \\ \delta \dot{r}_i \\ \delta \theta_i \end{bmatrix} = H_i\,\xi + V_i \tag{31}$$

The 8x1 state vector ξ is composed of changes in the elements δa_j and the bias components β_k.

$$\xi^T = (\delta a_1, \delta a_2, \delta a_3, \delta a_4, \beta_1, \beta_2, \beta_3, \beta_4).$$

The biases β_k could be direct biases in the measurements r_i, $\dot{r}_i$, θ_i or biases in tracking station locations $\bar{x}_i$, $\bar{y}_i$. V_i is a white noise vector with independent components. The matrix H_i is

$$H_i = \begin{bmatrix} \frac{\partial r_i}{\partial a_1} & \frac{\partial r_i}{\partial a_2} & \frac{\partial r_i}{\partial a_3} & \frac{\partial r_i}{\partial a_4} & \frac{\partial r_i}{\partial \beta_1} & \frac{\partial r_i}{\partial \beta_2} & \frac{\partial r_i}{\partial \beta_3} & \frac{\partial r_i}{\partial \beta_4} \\ \frac{\partial \dot{r}_i}{\partial a_1} & \frac{\partial \dot{r}_i}{\partial a_2} & \frac{\partial \dot{r}_i}{\partial a_3} & \frac{\partial \dot{r}_i}{\partial a_4} & \frac{\partial \dot{r}_i}{\partial \beta_1} & \frac{\partial \dot{r}_i}{\partial \beta_2} & \frac{\partial \dot{r}_i}{\partial \beta_3} & \frac{\partial \dot{r}_i}{\partial \beta_4} \\ \frac{\partial \theta_i}{\partial a_1} & \frac{\partial \theta_i}{\partial a_2} & \frac{\partial \theta_i}{\partial a_3} & \frac{\partial \theta_i}{\partial a_4} & \frac{\partial \theta_i}{\partial \beta_1} & \frac{\partial \theta_i}{\partial \beta_2} & \frac{\partial \theta_i}{\partial \beta_3} & \frac{\partial \theta_i}{\partial \beta_4} \end{bmatrix}$$

The differential equations of motion of the vector ξ are

$$\dot{\xi} = F\xi + Gu \tag{32}$$

where $F = [0]$

$$G^T = \begin{bmatrix} -(t-t_o) & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -(t-t_o) & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and $u^T = (f_x, f_y)$.

3.3 Estimation

A continuous-time Kalman filter is used for estimating the state vector ξ. The filter for the estimates $\hat{\xi}$ is

$$\dot{\hat{\xi}} = F\hat{\xi} + Gu + K(z - H\hat{\xi}) \tag{33}$$

where $z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ and $H = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix}$; with the subscripts designating the two trackers.

The matrix $K = PH^TR^{-1}$, where

$$P = E\,[ee^T],\quad e = \hat{\xi} - \xi,\quad R = \text{cov } V(t).$$

The equation for P is

$$\dot{P} = FP + PF^T + GQG^T - PH^TR^{-1}HP \tag{34}$$

where Q is the covariance matrix of the noise introduced by u; i.e., through the accelerating maneuvers. The matrix P is the covariance matrix of errors in the estimate. To do the error analysis, Equation (34) must be integrated forward from an assumed value P_o. The measurement covariance is defined as

$$M = E\left[(\hat{z}-\bar{z})(\hat{z}-\bar{z})^T\right]$$

where $\hat{z} = H\hat{\xi}$ and $\bar{z} = E(z)$. In terms of P,

$$M = HPH^T$$

M is also computed at each point along the trajectory, since its comparison with the covariance matrix R of the input data is another measure of the effectiveness of orbit refinement.

3.4 Simulation Involving Libration Point Trackers

For the error analysis the matrix P(t) is computed from an initial estimate P_o at $t = 0$. The value of P_o is determined by the uncertainties in the first (crude) orbit determination. Consider the refinement of the orbit elements as reflected in the error covariance matrix P at time $t > 0$.

Assume a probe trajectory heading away from Earth at a 135-degree angle with the +x axis (Figure 143). This resembles a typical escape trajectory for a space probe

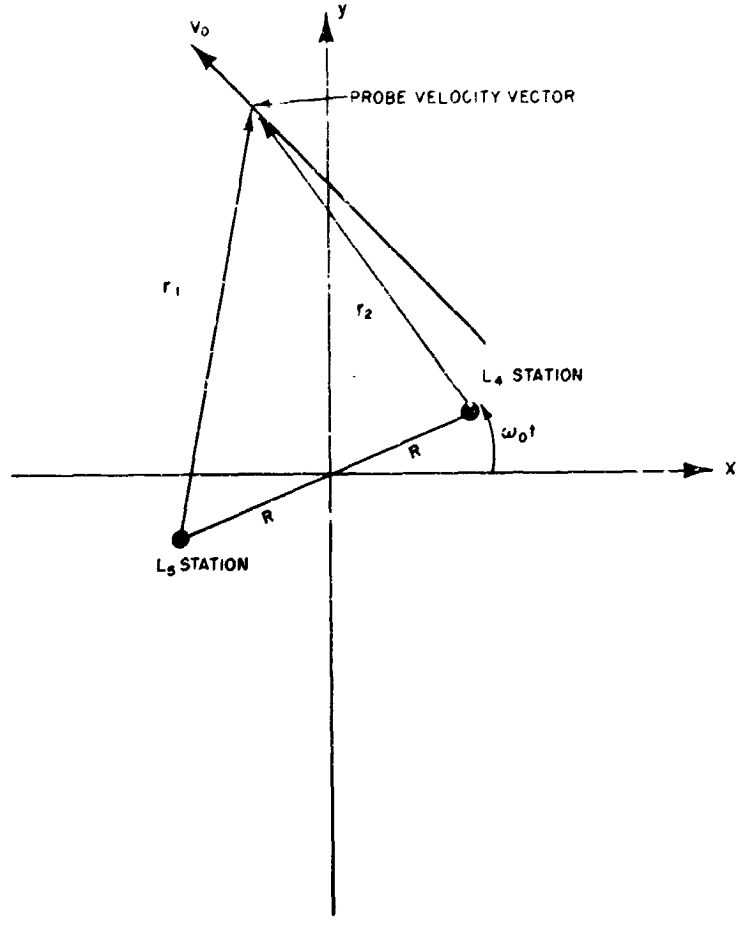


Figure 143. Geometry of Observations

heading toward Mars. The model is quite insensitive to the trajectory angle, since it turns out that errors along the range vector are refined to much the same degree regardless of the direction of $\bar{V}_\infty$. Cross-range errors are likewise essentially invariant with the trajectory angle.

It is assumed that typical rms tracking errors are 1.0 km in range, 3 mm/s in range rate, and 10^{-5} radians ($\simeq 2$ seconds of arc) in angle. These values are consistent with those selected in Section 2 after the best current claims are downgraded in recognition of astrophysical uncertainties (though σ_γ is still slightly optimistic). The cases of range-range rate data and range-range rate-angle data will be treated.

Let the station locations of the libration point satellites at L_4 and L_5 be

$$\begin{aligned}
\bar{x}_1 &= R\left[\cos\left(\omega_o t + \tfrac{5\pi}{6}\right) + 0.2\cos\left(0.3\omega_o t\right)\right] \\
\bar{y}_1 &= R\left[\sin\left(\omega_o t + \tfrac{5\pi}{6}\right) + 0.2\sin\left(0.3\omega_o t\right)\right] \\
\bar{x}_2 &= R\left[\cos\left(\omega_o t + \tfrac{\pi}{6}\right) + 0.2\cos\left(0.3\omega_o t\right)\right] \\
\bar{y}_2 &= R\left[\sin\left(\omega_o t + \tfrac{\pi}{6}\right) + 0.2\sin\left(0.3\omega_o t\right)\right]
\end{aligned} \tag{35}$$

The errors due to biases β_k in the elements of the libration point satellites are

$$\begin{aligned}
\delta\bar{x}_i &= \beta_i \cos\omega_o t - \beta_{i+2}\sin\omega_o t \\
\delta\bar{y}_i &= \beta_i \sin\omega_o t + \beta_{i+2}\cos\omega_o t \\
\delta\dot{\bar{x}}_i &= -\omega_o\left[\beta_i \sin\omega_o t + \beta_{i+2}\cos\omega_o t\right] \\
\delta\dot{\bar{y}}_i &= \omega_o\left[\beta_i \cos\omega_o t - \beta_{i+2}\sin\omega_o t\right]
\end{aligned} \tag{36}$$

where i = 1,2. The functional forms of these errors reflect the fundamental mode of motion about a libration point, which shall serve as a simplified dynamic model. It turns out to be an adequate approximation of the motion for purposes of these orbit determination studies. The biases are β_1, β_3 for tracker one and β_2, β_4 for tracker two. The initial estimates of rms errors in β_k are assumed to be 0.1 km. It is also assumed that $R = 3.0 \times 10^5$ km and $\omega_o = 2\pi/27$ radians/day. Finally, the case R = 0, representing a tracking configuration with stations on the earth, will be compared with the libration point trackers.

The nominal elements of the space probe are taken to be

$$a_1 = 0$$

$$a_2 = 10^8\,\text{km}\,(10^7\,\text{km})$$

$$a_3 = -\frac{1.73\times10^6}{\sqrt{2}}\ \frac{\text{km}}{\text{day}}\left(-\frac{0.173\times10^6}{\sqrt{2}}\ \frac{\text{km}}{\text{day}}\right) \tag{37}$$

$$a_4 = \frac{1.73\times10^6}{\sqrt{2}}\ \frac{\text{km}}{\text{day}}\left(\frac{0.173\times10^6}{\sqrt{2}}\ \frac{\text{km}}{\text{day}}\right)$$

which corresponds to the configuration in Figure 143. Note that 10^7 km, a tenth of an astronomical unit (the value for a_2 in parentheses), should be typical of the ranges prevailing during the early tracking operations for a Martian transfer trajectory. On the other hand, 10^8 km represents a distance that is well into a Martian flyby mission and was also used in these calculations by way of comparison.

3.5 Numerical Results

Numerical test runs were conducted for geocentric distances to the space vehicle in the order of 10^7 and 10^8 km. The geocentric radius to a libration point was taken as $R = 3.0 \times 10^5$ km, thus providing a base line of 6.0×10^5 km. To examine the effects of trilateration from this base line, the option R = 0 was also exercised, which represents the case of terrestrial trackers. The possible data combinations were restricted to the usual two: range-range rate and range-range rate-angles. Besides the rms angle error of 10^{-5} radians, more refined measurements of $\sqrt{10 \times 10^{-6}}$ and 10^{-5} radians were included as an extrapolation to the future. The data rate was assumed as one measurement per day in every case. (In view of the long flight times and hence large numbers of observations involved, this still justifies the use of a continuous model for the estimation process.)

Nearly all possible combinations of the above options were run, though only the most significant ones will be discussed. With range-range rate data only, a near-linear escape trajectory offers little opportunity for improvement of the cross-range errors in the absence of trilateration from a long base line. This is mainly due to the fact that the tracking geometry undergoes little change with time. The effect is borne out by Runs 1 and 2 of Table 62 and the top curve in Figure 144. This situation improves significantly if the distance between libration points is used as a tracking base line, especially at nominal ranges of the order 10^7 km (Runs 3 and 4; bottom curve on Figure 144). Whether or not range or angle measurements are used to take advantage of the base line seems to be relatively unimportant; i.e., the

tracking performances based on trilateration and triangulation are identical for practical purposes. However, optical angles do represent an asset even without a base line for moderate tracking distances, as shown by Run 6 of Table 62 and the middle curve on Figure 144. Increasing the accuracy of optical data beyond 2 s of the arc does relatively little to enhance absolute position accuracies but eliminates the need for a base line, as demonstrated by Runs 13 and 14 of Table 62. This means that tracking relays could, for instance, be placed in synchronous orbits if optical data with rms errors of 0.2 s were achievable. On the other hand, if trilateration were the governing mechanism of orbit determination, such a reduction of baseline would cut the effective tracking distances to about 10^6 km for the same σ in vehicle position.

Finally, Table 63 exhibits the effect of using relatively low-grade range data, which might for instance reflect a deterioration of the ephemeris for the orbital tracking stations. It is interesting to note that this influences only the down-range position errors in the numerical runs, whereas the interplay of various effects in the cross-range errors remains essentially unchanged.

7.6 Conclusions

The numerical results of this study show that state-of-the-art optics, using stellar references (σ = 2 s), and DSIF-type range or range-rate data (effective σ_r = 1 km, $\sigma_{\dot{r}}$ = 3 mm/sec) are roughly equivalent as tracking data for distances up to 10^7 km from the earth. This comparison is of course especially valid for lunar missions. While range tracking requires a base line extending to the libration points for vehicle position errors of the order of 1 km, the same can be accomplished with optics from near-Earth satellites.

The expressions in Equation (36) for tracker location errors are, admittedly, a simplification. Not much refinement of the β_k was experienced in the numerical simulations, and the values assumed for σ_r reflect some conservatism in this respect. A further strengthening of confidence in this kind of study, with a possible improvement of trajectory accuracies in the near-Earth phase, will require a refinement of the ephemerides for orbital tracking stations. This implies more elaborate models for the librational motion. As noted earlier, these models suggest a

Table 62

ACCURACIES AT 30 DAYS – ONE MEASUREMENT/DAY
RMS range = 1 km
RMS range rate = 0.3 cm/s

Run No.	Type of Data	RMS Angle (μrad)	Nominal Tracking Distance (km)	RMS Bias (km)	Libration Point Radius=R (km)	Position Error (km)		Velocity Error (cm/s)	
						Cross Range (x)	Down Range (y)	Cross Range ($\dot{x}$)	Down Range ($\dot{y}$)
1	r+$\dot{r}$	–	10^8	0.1	= 0	98.6	0.179	1.17	1.16
2	r+$\dot{r}$	–	10^7	0.1	= 0	98.3	0.179	1.17	1.16
3	r+$\dot{r}$	–	10^8	0.1	$3X10^5$	43.0	0.179	0.511	0.508
4	r+$\dot{r}$	–	10^7	0.1	$3X10^5$	4.74	0.172	0.120	0.058
5	r+$\dot{r}$+A	10	10^8	0.1	= 0	78.4	0.179	0.933	0.926
6	r+$\dot{r}$+A	10	10^7	0.1	= 0	12.8	0.179	0.194	0.152
7	r+$\dot{r}$+A	10	10^8	0.1	$3X10^5$	40.8	0.179	0.486	0.482
8	r+$\dot{r}$+A	10	10^7	0.1	$3X10^5$	4.44	0.172	0.118	0.054
9	r+$\dot{r}$+A	$\sqrt{10}$	10^8	0.1	= 0	37.7	0.179	0.462	0.446
10	r+$\dot{r}$+A	$\sqrt{10}$	10^7	0.1	= 0	4.08	0.177	0.128	0.050
11	r+$\dot{r}$+A	$\sqrt{10}$	10^8	0.1	$3X10^5$	29.6	0.179	0.363	0.350
12	r+$\dot{r}$+A	$\sqrt{10}$	10^7	0.1	$3X10^5$	3.08	0.170	0.112	0.039
13	r+$\dot{r}$+A	1.0	10^8	0.1	= 0	12.8	0.179	0.194	0.152
14	r+$\dot{r}$+A	1.0	10^7	0.1	= 0	1.29	0.164	0.096	0.021
15	r+$\dot{r}$+A	1.0	10^8	0.1	$3X10^5$	12.4	0.178	0.187	0.147
16	r+$\dot{r}$+A	1.0	10^7	0.1	$3X10^5$	1.24	0.160	0.089	0.020

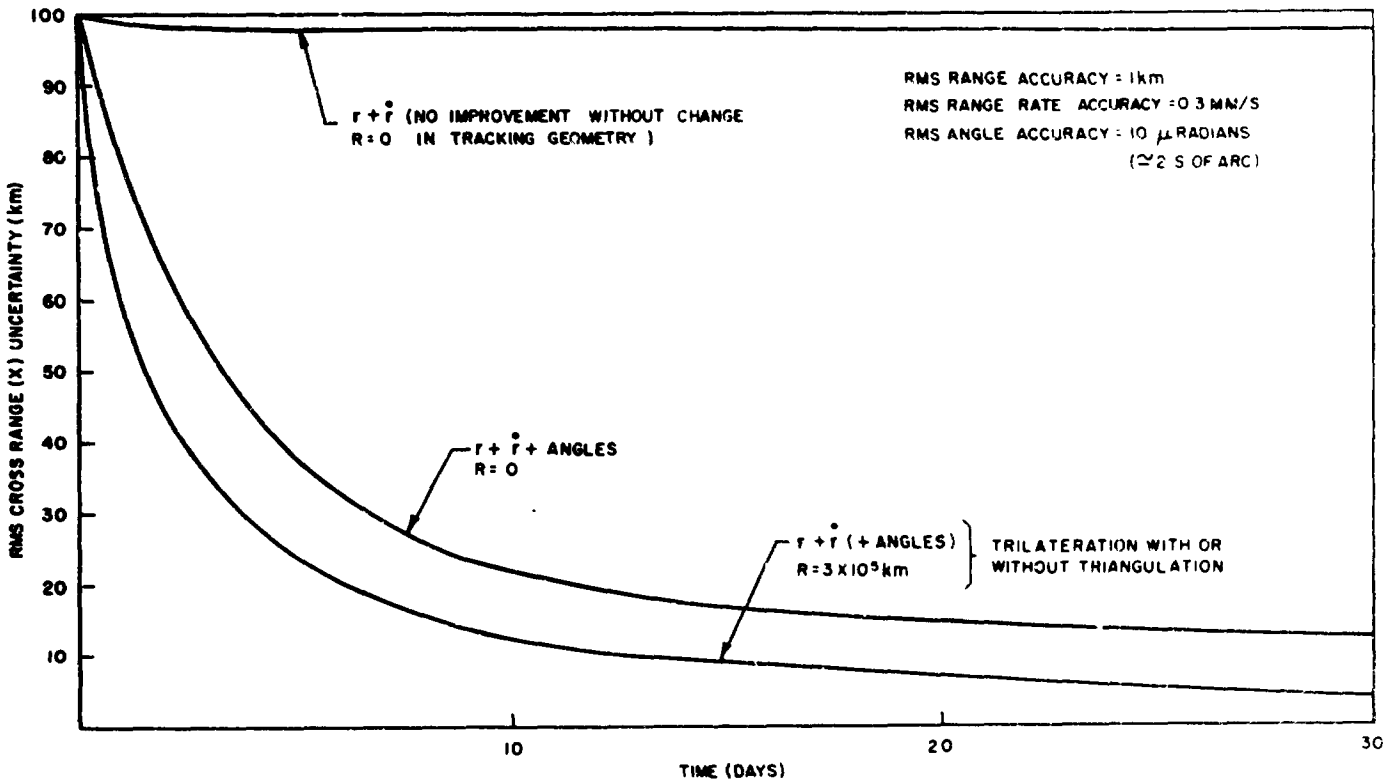


Figure 144. RMS Cross-Range Position Uncertainty Vs. Time at Distances of About 10^7 km

Table 63

ACCURACIES AT 30 DAYS ONE MEASUREMENT/DAY
RMS range = 100 km
RMS range rate = 0.3 cm/s

Run No.	Type of Data	RMS Angle (μrad)	Nominal Tracking Distance (km)	RMS Bias (km)	Libration Point Radius=R (km)	Position Error (km)		Velocity Error (cm/s)	
						Cross Range (x)	Down Range (y)	Cross Range (x)	Down Range (y)
1	r+$\dot{r}$	–	10^8	0.1	= 0	98.6	12.8	1.18	1.17
2	r+$\dot{r}$	–	10^7	0.1	= 0	98.3	12.8	1.18	1.16
3	r+$\dot{r}$	–	10^8	0.1	$3X10^5$	60.8	12.8	0.737	0.724
4	r+$\dot{r}$	–	10^7	0.1	$3X10^5$	7.66	12.7	0.205	0.104
5	r+$\dot{r}$+A	10	10^8	0.1	= 0	78.3	12.8	0.940	0.925
6	r+$\dot{r}$+A	10	10^7	0.1	= 0	12.8	12.8	0.251	0.156
7	r+$\dot{r}$+A	10	10^8	0.1	$3X10^5$	55.0	12.8	0.670	0.655
8	r+$\dot{r}$+A	10	10^7	0.1	$3X10^5$	6.55	12.7	0.200	0.092
9	r+$\dot{r}$+A	$\sqrt{10}$	10^8	0.1	= 0	37.7	12.8	0.489	0.447
10	r+$\dot{r}$+A	$\sqrt{10}$	10^7	0.1	= 0	4.09	12.6	0.200	0.062
11	r+$\dot{r}$+A	$\sqrt{10}$	10^8	0.1	$3X10^5$	33.9	12.8	0.443	0.405
12	r+$\dot{r}$+A	$\sqrt{10}$	10^7	0.1	$3X10^5$	3.59	12.6	0.186	0.060
13	r+$\dot{r}$+A	1.0	10^8	0.1	= 0	12.8	12.8	0.251	0.156
14	r+$\dot{r}$+A	1.0	10^7	0.1	= 0	1.30	12.1	0.167	0.042
15	r+$\dot{r}$+A	1.0	10^8	0.1	$3X10^5$	2.6	12.8	0.248	0.155
16	r+$\dot{r}$+A	1.0	10^7	0.1	$3X10^5$	1.28	11.9	0.161	0.042

functional form of the residuals in the ephemerides of tracking relays which is slowly varying rather than a constant bias or pure noise. The proper treatment of such correlated errors would be the most important statistical feature of a follow-on effort. The ultimate limits on tracking accuracy from orbital relays will derive from geophysical uncertainties and occulation conditions.

In connecting the near-Earth phase to a heliocentric representation of the trajectory, the problem of units discussed in Section 2 was encountered. While a trilateration procedure implies terrestrial measurements (i.e., the meter) as a unit of length, the interplanetary phase uses ranges in terms of light seconds. Thus, the speed of light becomes one of the quantities undergoing refinement.

Finally, a three-dimensional extension of the present work should be considered for complete generality. To first order, the dynamic perturbations and trajectory estimation errors normal to the ecliptic are uncoupled from the two-dimensional model. If optical data are relied upon, the same set of near-Earth trackers will also cope with the three-dimensional situation. If trilateration is to be simulated in the third dimension, one would have to propose tracking relays in highly eccentric earth orbits whose major axes are essentially normal to the ecliptic.

4. NAVIGATION NEAR MARS

Consider a Mars flyby trajectory in a planet-centered hyperbolic orbit (Figure 145). The flyby vehicle F is observed from a planet-centered satellite S by means of range, range rate, and (possibly) angle measurements. An Earth-centered range, range rate, and (possibly) angle tracker E observes the satellite and the spacecraft, but is limited by long range and poor geometry. The observer S would presumably improve the accuracy of orbit refinement and speed of error response. The question to be answered by the present error analysis is: How much is the performance improved by addition of observer S?

4.1 Orbit Geometry and Observables

It is assumed that for most of the salient features of the problem a planar solution will suffice. It is also assumed

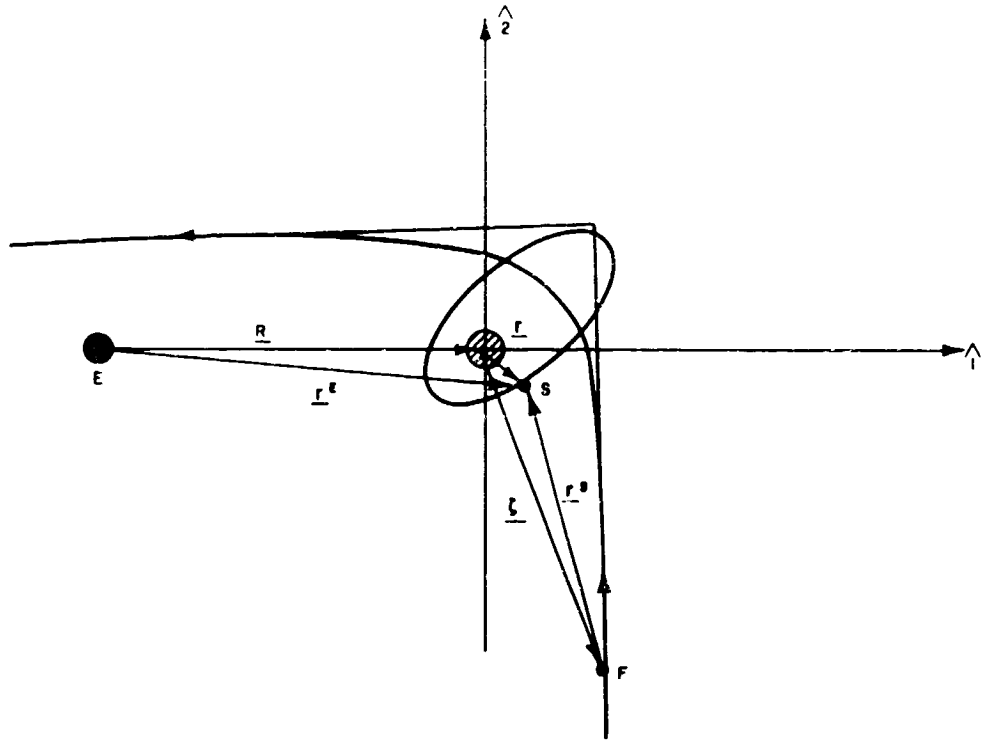


Figure 145. Geometry of Flyby (F) Trajectory and Satellite (S) Orbit, Mars at Origin of Coordinates

SATELLITE PARAMETERS:

a_E = 4400 km
$\mathcal{L}$ = 0.1
m = 0
τ_E = 0

FLYBY PARAMETERS:

a_H = 1720 km
ω = 0
θ = 5.53
τ_H = 16.37 HR

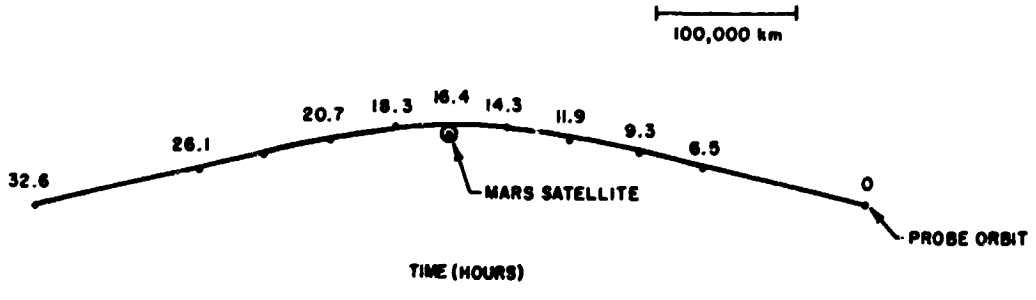


Figure 146. Flyby Trajectory

that the problem time is short enough to allow the Earth-planet line to be fixed in inertial space. It is denoted by unit vector $\hat{1}$. Unit vector $\hat{2}$ is taken perpendicular to $\hat{1}$ and through the planet center as in Figure 145. The position of E is denoted by vector $\underline{\rho}$, which is assumed to be a function of time, and four orbit elements a_1, a_2, a_3, a_4. The position of S is denoted by a vector $\underline{r}$ which is assumed to be a function of time, and elements b_1, b_2, b_3, b_4. The E-planet distance vector is $\underline{R}$. The derived vectors $\underline{r}^E$ from E to S and $\underline{r}^S$ from E to S are then

$$\begin{aligned} \underline{r}^E &= \underline{R} + \underline{r} \\ \underline{r}^S &= \underline{r} - \underline{\rho} \end{aligned} \tag{38}$$

The observed variables are denoted λ_i $(i = 1,2,\ldots,10)$:

$$\lambda_1 = |\underline{r}^E| = \left[\underline{R}^2 + 2\underline{r}\cdot\underline{R} + \underline{r}^2\right]^{1/2} \tag{39}$$

$$\lambda_2 = \frac{\dot{\underline{r}}^E \cdot \underline{r}^E}{|\underline{r}^E|}$$

$$\lambda_3 = \sin^{-1}\left[\underline{r}^E \cdot \hat{2} / |\underline{r}^E|\right]$$

$$\lambda_4 = |\underline{r}^S| = \left[\underline{r}^2 - 2\underline{r}\cdot\underline{\rho} + \underline{\rho}^2\right]^{1/2}$$

$$\lambda_5 = \frac{\dot{\underline{r}}^S \cdot \underline{r}^S}{|\underline{r}^S|}$$

$$\sin\lambda_6 = \left(\underline{r}^S \cdot \hat{2} / |\underline{r}^S|\right)$$

$$\cos\lambda_6 = \left(\underline{r}^S \cdot \hat{1} / |\underline{r}^S|\right)$$

$$\lambda_7 = |\underline{r}| = \left(\underline{r}\cdot\underline{r}\right)^{1/2}$$

$$\lambda_8 = |\underline{R}+\underline{\rho}|$$

$$\lambda_9 = \dot{\underline{\rho}} \cdot (\underline{\rho}+\underline{R}) / |\underline{\rho}+\underline{R}|$$

$$\lambda_{10} = \sin^{-1}\left[(\underline{\rho}+\underline{R})\cdot\hat{2} / |\underline{\rho}+\underline{R}|\right]$$

Here $\lambda_1, \lambda_2, \lambda_3$ are range, range rate, and angle of the satellite relative to Earth. $\lambda_4, \lambda_5, \lambda_6$ are range, range rate, and angle of the satellite relative to the probe. λ_7 is the altitude of the satellite above Mars. $\lambda_8, \lambda_9, \lambda_{10}$ are range, range rate, and angle of the probe relative to Earth.

A vector $\alpha^T = (a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4)$ of the hyperbolic and elliptic orbit elements must be estimated. Let α represent a nominal trajectory and the deviations from nominal $\delta\alpha = \alpha - \tilde{\alpha}$. The sensitivity matrix h^i, relating errors in elements to errors in measurements $\delta\lambda_i$, is required. Thus $h^i = \partial\lambda_i/\partial\tilde{\alpha}$ defines a row vector of partial derivatives

$$h^i = \left(\frac{\partial\lambda_i}{\partial\tilde{\alpha}_1}, \ldots, \frac{\partial\lambda_i}{\partial\tilde{\alpha}_8}\right)$$

This yields a vector z_i of observations

$$z_i = h^i\delta\alpha + V_i = \delta\lambda_i$$

where V_i is a white noise representing observation error. The covariance of V_i is $E[V_iV_j] = R_{ij}\delta_{ij}$.

4.2 Kalman Filter[10]

Measurements are taken at equal intervals in the true anomaly of the probe orbit. The optimal estimate is updated at each measurement point. The estimate of $\delta\alpha$ at instant n is

$$\hat{\delta\alpha}_{n+1} = \hat{\delta\alpha}_n + K_n\left(z_n - H_n\hat{\delta\alpha}_n\right) \tag{40}$$

where $H_n^T = \left(h_n^{1T}, h_n^{2T}, \ldots, h_n^{10T}\right)$ is the sensitivity matrix of the combined observations and z_n is the combined measurement vector at the n^{th} sample point. The 8×10 matrix $K_n = P_nH_n^TR_n^{-1}$, when P_n is the covariance matrix of orbit elements, weights present observations by the usual rationale: Observations believed to be very accurate compared with current orbit element estimates will be heavily weighted, whereas less accurate observations will have little effect on subsequent estimates.

The state covariance matrix is the solution to the equation

$$P_{n+1} = P_n - P_nH_n^T\left(R_n + H_nP_nH_n^T\right)^{-1}H_nP_n \tag{41}$$

The matrix $H_n^T\left(R_n+H_nP_nH_n^T\right)^{-1}H_n$ is called the information rate matrix and has the effect of reducing the diagonal elements of P_n. This equation yields a complete error analysis of the orbit determination problem. P_n (in double precision) is solved recursively, given its initial value P_o. These iterative computations occur at equal increments of f, the true anomaly of the probe. The given value of f is taken and F, t are computed by known formulas. Thus the hyperbolic orbit and its partial derivatives may be computed. Using t, the elliptic Kepler equation is inverted; the position in the elliptic orbit and its partial derivatives are then computed (see Appendix 11).

The vectors h^i are computed using the formula

$$h^i = (J_{Ai}A, J_{Bi}B)$$

where the matrices are defined as follows with the hyperbolic and elliptic elements given by $a_1 = a_h$, $a_2 = e$, $a_3 = \omega$, $a_4 = \tau$, $b_1 = a_t$, $b_2 = \gamma$, $b_3 = m$, $b_4 = T$

$$A = \begin{bmatrix} \frac{\partial \rho_1}{\partial a_1} & \frac{\partial \rho_1}{\partial a_2} & \frac{\partial \rho_1}{\partial a_3} & \frac{\partial \rho_1}{\partial a_4} \\ \frac{\partial \rho_2}{\partial a_1} & \frac{\partial \rho_2}{\partial a_2} & \frac{\partial \rho_2}{\partial a_3} & \frac{\partial \rho_2}{\partial a_4} \\ \frac{\partial \rho_3}{\partial a_1} & \frac{\partial \rho_3}{\partial a_2} & \frac{\partial \rho_3}{\partial a_3} & \frac{\partial \rho_3}{\partial a_4} \\ \frac{\partial \rho_4}{\partial a_1} & \frac{\partial \rho_4}{\partial a_2} & \frac{\partial \rho_4}{\partial a_3} & \frac{\partial \rho_4}{\partial a_4} \end{bmatrix}$$

$$B = \begin{bmatrix} \frac{\partial r_1}{\partial b_1} & \frac{\partial r_1}{\partial b_2} & \frac{\partial r_1}{\partial b_3} & \frac{\partial r_1}{\partial b_4} \\ \frac{\partial r_2}{\partial b_1} & \frac{\partial r_2}{\partial b_2} & \frac{\partial r_2}{\partial b_3} & \frac{\partial r_2}{\partial b_4} \\ \frac{\partial r_3}{\partial b_1} & \frac{\partial r_3}{\partial b_2} & \frac{\partial r_3}{\partial b_3} & \frac{\partial r_3}{\partial b_4} \\ \frac{\partial r_4}{\partial b_1} & \frac{\partial r_4}{\partial b_2} & \frac{\partial r_4}{\partial b_3} & \frac{\partial r_4}{\partial b_4} \end{bmatrix}$$

$$J_{Ai} = \begin{bmatrix} \frac{\partial \lambda_i}{\partial \rho_1}, & \frac{\partial \lambda_i}{\partial \rho_2}, & \frac{\partial \lambda_i}{\partial \rho_3}, & \frac{\partial \lambda_i}{\partial \rho_4} \end{bmatrix}$$

$$J_{Bi} = \begin{bmatrix} \frac{\partial \lambda_i}{\partial r_1}, & \frac{\partial \lambda_i}{\partial r_2}, & \frac{\partial \lambda_i}{\partial r_3}, & \frac{\partial \lambda_i}{\partial r_4} \end{bmatrix}$$

The x and y components of vectors are denoted $\rho_1 = \underline{\rho} \cdot \hat{1}$, $\rho_2 = \underline{\rho} \cdot \hat{2}$, etc.

A new state vector will now be defined in rectangular variables as $X^T = (\rho_1, \rho_2, \dot{\rho}_1, \dot{\rho}_2, r_1, r_2, \dot{r}_1, \dot{r}_2) = (\rho_1, \rho_2, \rho_3, \rho_4, r_1, r_2, r_3, r_4)$. Define $M = E(\delta X \delta X^T)$ as the error covariance matrix in rectangular coordinates. The relation $M_n = SP_nS^T$ can be shown to hold, where

$$S = \begin{bmatrix} A & O \\ O & B \end{bmatrix}$$

By starting with a value M_o and setting $P_o = S^{-1}M_oS^{T-1}$ recursive computation of P_n can begin. At the output P_n is converted into M_n by the above formula.

4.3 Numerical Results

Several cases were run on the digital computer. Equation (41) was solved for P_n, assuming steps of $f_n - f_{n-1} = 0.02$ radian in the true anomaly. The input covariance matrix M_o was chosen consistent with the programs for interplanetary transfer described in Section 2. In fact, the rms values of initial uncertainties* were

Probe	Satellite
Position 5.1 km	1.41 km
Velocity 1.16 cm/s	24.1 cm/s

The rms noise in the measurements was assumed† to be:

	From Earth	From Satellite
Range	1 km	0.3 km
Range Rate	1 cm/s	10 cm/s

The nominal orbits for the space probe and the Mars orbiter are shown in Figure 146. The following cases were considered, using the above conditions: (1) probe observed from Earth, (2) probe observed from satellite, and (3) combined observations from (1) and (2). As indicated above, the initial uncertainties in the satellite state vector were assumed to result from earth tracking.

Figures 147a and 147b show the evolution of position uncertainty of the probe and of the satellite over a critical period of about 15 hours. For the probe the Mars orbiter provides a great advantage in response time over Earth-based observations. This effect brings the position error from about 5 km to about 1.2 km during an interval of 0.2 radians in true anomaly or 9 hours of tracking time. In any case the asymptotic position error is less than 0.5 km (after 157 hours of tracking).

Figure 148 shows the rms error in the probe position vs. its true anomaly subsequent to a sudden rise in position uncertainty represented by $\sigma_p = 3$ km. This might be due to accidental interruption of the tracking operation, loss of

*The value of σ_r for the probe is somewhat conservative if compared with the results obtainable by the program of Section 2. Similarly, the orbit of a Mars satellite may be determined to somewhat better accuracy from the Earth if tracked relative to the planet.

†For σ_r and $\sigma_{\dot{r}}$ from the Earth, published values for DSIF were degraded, as in Section 2, to allow for astrophysical uncertainties and position errors of the near-Earth trackers themselves. From the Mars orbiter to probe, the signal to noise ratio (compared to DSIF) would be lowered because the vehicle antenna is smaller than a ground antenna and raised because the range is about 0.001 of interplanetary distances; the two effects cancel and thus SNR is preserved. The effect of timing uncertainties for Doppler recordings is crucial. A space-born crystal oscillator can be stabilized (short term) to 1 part in 10^{11}. This yields a Doppler error of 3 cm/s. In this study rms accuracies of 0.3 km and 10 cm/s were assumed as conservative estimates. It is also pessimistic to assume no improvement as the range between Mars orbiter and space probe decreases.

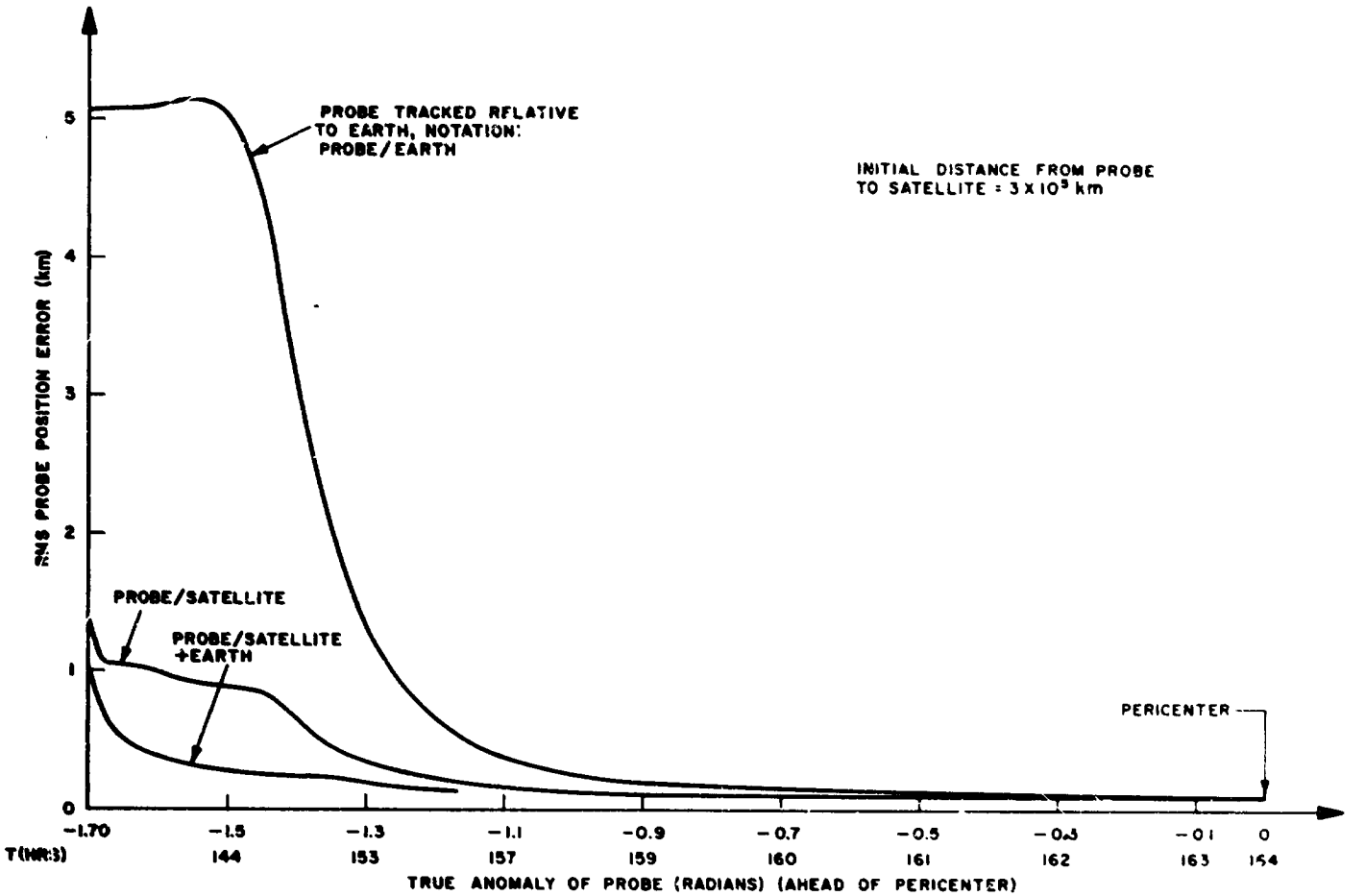


Figure 147a. Plot of probe position error vs. true anomaly

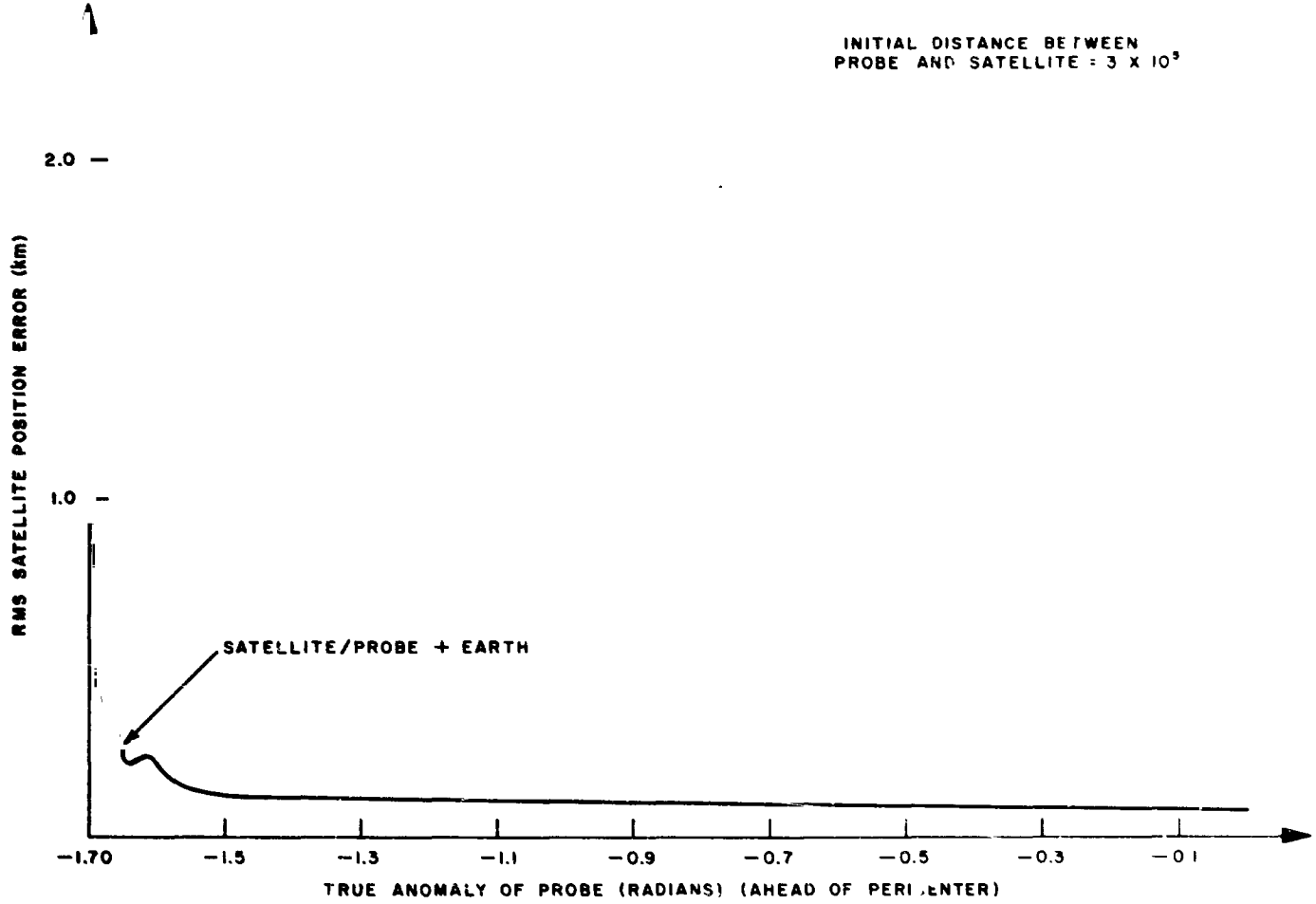


Figure 147b. Plot of satellite position error vs. true anomaly

tracking information, or other anomalies of the mission. Again, observations from the Mars orbiter contribute significantly toward accelerated reacquisition during this transient period. Note that this event was chosen to occur rather close in time to perigee.

Figure 149 shows a plot similar to Figure 148, giving the rms velocity error in response to a transient uncertainty of 0.8 m/s. This could represent the recovery from a midcourse correction with loss of communication. Again, the tracking assistance from the Mars orbiter is felt in a way similar to that of Figure 148.

4.4 Conclusions

The most significant aspect of the flyby maneuver is the quick-response capability offered by the Mars orbiter for absorbing sudden uncertainties in the space probe ephemeris. This may be particularly useful for checking the probe position along its trajectory, as needed for the precise dispatch of automated excursion vehicles to the planet surface. The value of the Mars orbiter as a navigation aid during flyby is due to the prior orbit refinement it receives from the Earth. Its very presence as a tracking aid during the flyby maneuver is the governing feature of this particular situation. In this respect, the addition of optical angle measurements, with $\sigma = 2$ s, to the range and range-rate data generated between probe and orbiter is of secondary importance. This was corroborated by a few exploratory results not included among the figures given here. The same will probably hold for three-way Doppler measurements between Earth, space probe, and Mars orbiter. It seems particularly important to explain the effectiveness with which the Martian satellite orbit may be determined from Earth. Bias errors entering through terrestrial trackers or geophysical uncertainties would seem to make such a tracking operation rather questionable if it were not for the fact that differential observations of the orbiter relative to Mars can eliminate their effect. A major residual uncertainty would then be connected with the timing of the orbiter's motion, which translates into an in-track error. This can stand a more detailed study.

For the three-dimensional case, the out-of-ecliptic errors of the space probe may be reduced by using a highly inclined Mars orbiter at considerable altitude. This aspect of the three-dimensional problem may yield very interesting results. Similarly, the proper treatment of bias errors resulting from the tracking instruments and astrophysical constants is important if a complete understanding of terminal navigation is to be achieved. Such studies would be especially important if one were to investigate navigation problems connected with the landing and return rendezvous of a surface excursion vehicle or manned, round-trip flyby missions.

5. INTRAGALACTIC NAVIGATION

An intragalactic transfer; i.e., a probe sent from the Earth to the outer reaches of the solar system and beyond, will differ from the problems of Section 2 in two ways. The trip will be longer, and for a large portion of the flight the probe will be influenced mainly by solar gravity. Nevertheless, there is sufficient similarity for this situation to be considered a scaled-up analog of the interplanetary case.

Such a mission is simulated and the results described briefly in this section. As in Section 2, the effect of the inclusion of highly accurate angular information upon the position and velocity uncertainties of the spacecraft are of primary interest.

Numerical Results and Conclusions

The results of the near-Earth phase of the mission; i.e., the first 10 days, will be carried over without change; they are taken from Table 59 and Figures 135a and 135b. Thus the position and velocity uncertainties of the probe are as obtained previously. From Day 10 on, however, the probe will be acted upon by the gravities of the Earth and the Sun only; i.e., the Martian gravity is omitted from the computer program. The data are processed optimally,* using the tracking parameters given in Table 60, part IIIa. The position and velocity uncertainties of the probe are denoted σ_p and σ_v respectively, and they are illustrated for data processing with (r, $\dot{r}$) and (r, $\dot{r}$, angle)-type observations in Figures 150a and 150b for the first 450 days of the mission.

These curves are to be compared with Figures 137a and 137b, where a 100-day interval of an interplanetary mission with similar tracking parameters was considered. As expected, the results are similar for the first 100 days of the flight. That is, the effect of highly accurate angular data, such as would be obtained from exoatmospheric optical instrumentation, is to reduce position and velocity uncertainties significantly in the early portion (say, the first two months) of an intragalactic flight. There is a gradual but continual increase in the uncertainties of position and velocity as the time of flight becomes very large. This might be expected from the increasing Earth-probe distance, since the triangulation effect of the Earth's orbital movement relative to the probe becomes less pronounced with increasing range. (The drop in position uncertainty after Day 400, which seems to fall out of line, corresponds to a significant, but periodic decrease in the Earth-probe distance in this region, which is caused by the Earth's annual motion.) One could inquire if a regular trilateration

*That is, using Equation (15).

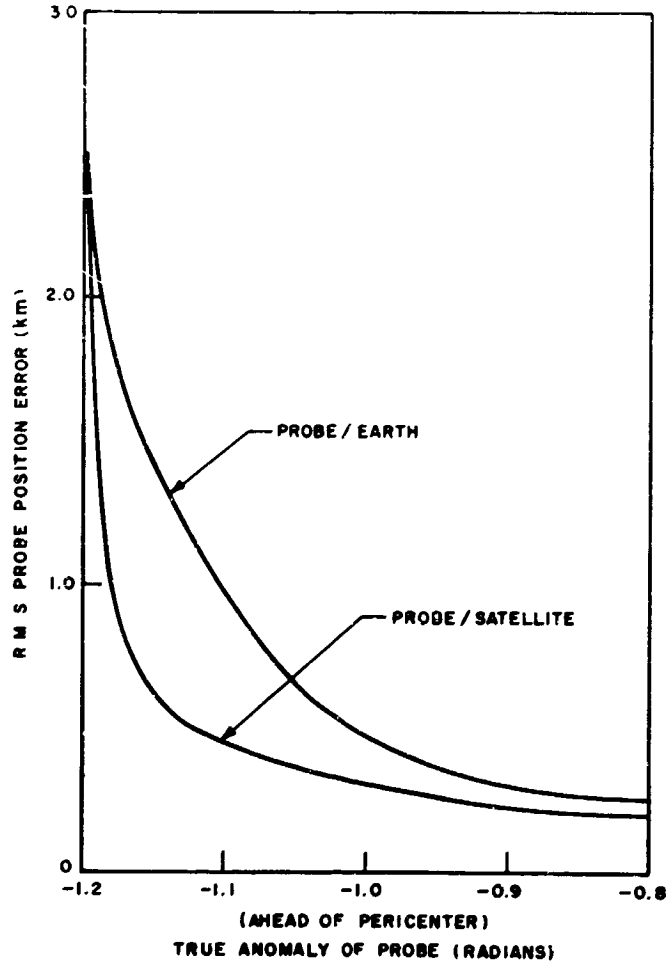


Figure 148. Effect of transient position uncertainty

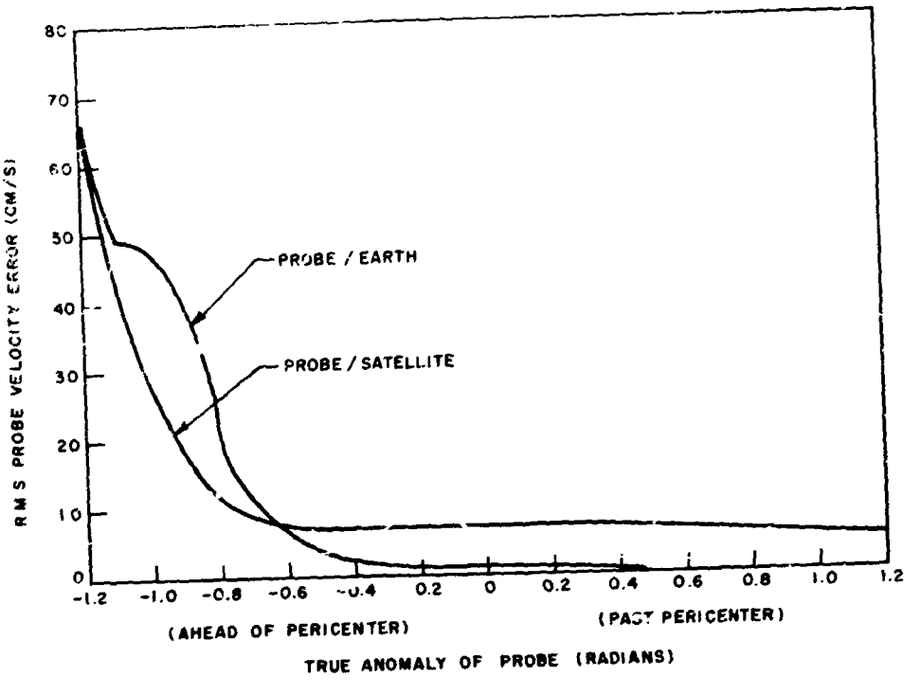


Figure 149. Effect of a transient velocity uncertainty

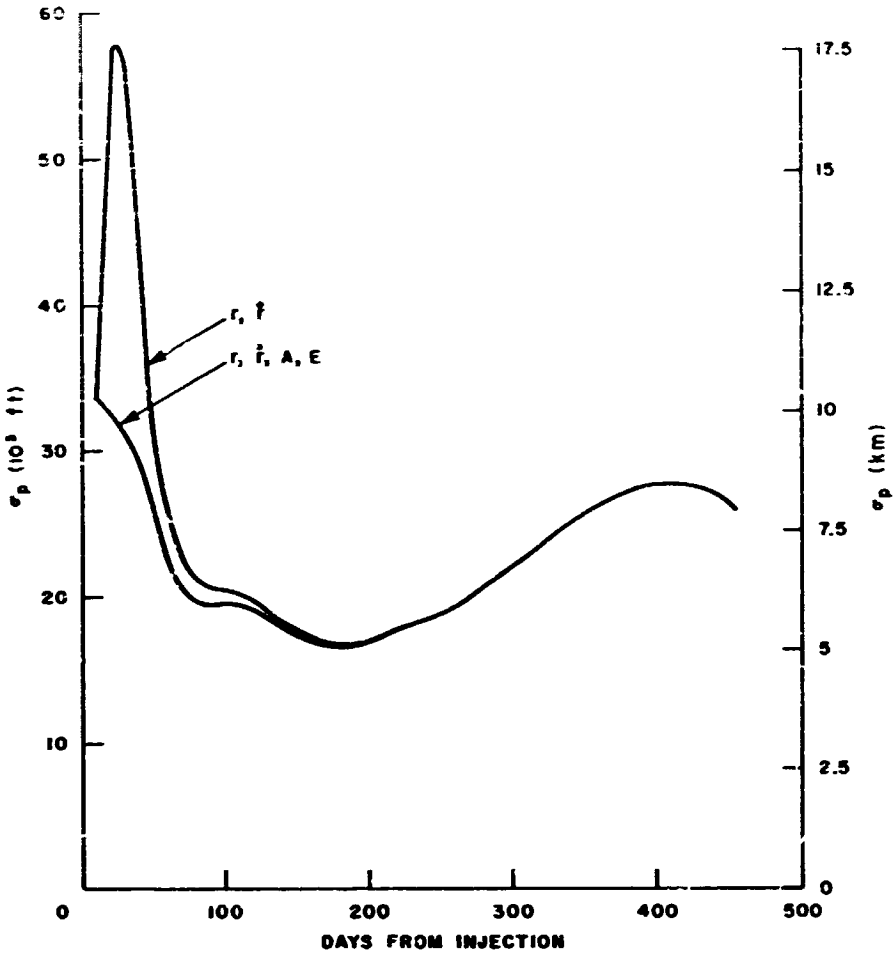


Figure 150a. Effect of angles – intragalactic transfer

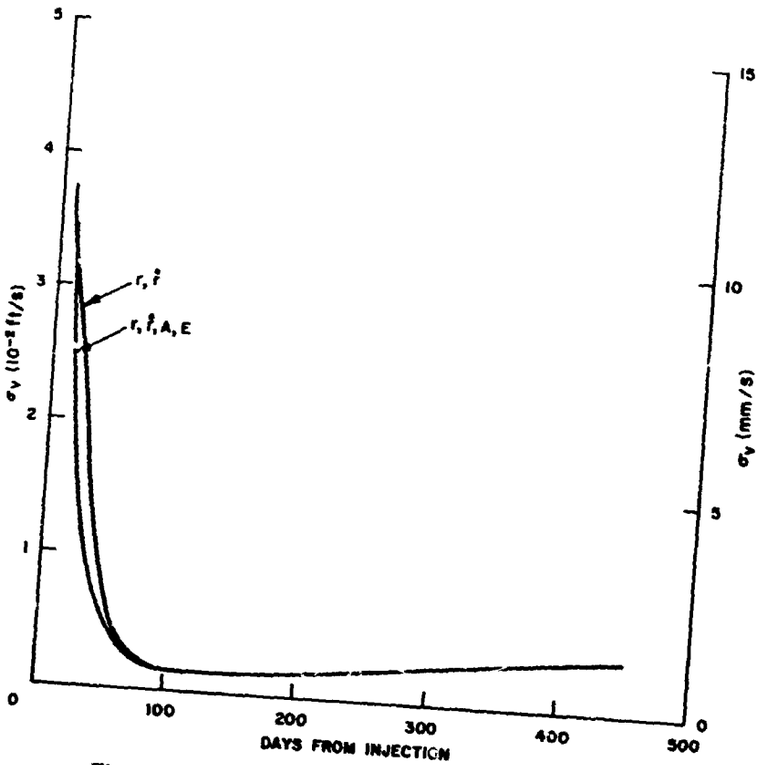


Figure 150b. Effect of angles–intragalactic transfer

scheme, extended over interplanetary distances, might not prove advantageous. Thus, for example, a Mars or Venus orbiter and near-Earth satellites could supply the necessary base line. Again, the payoffs from this scheme are highly dependent on errors in this base line, which means that adequate models for ephemeris errors and bias effects must be provided.

6. SUMMARY AND PENDING PROBLEMS

The merits of optical angle measurements, if added to range and range-rate data, consist of rapid trajectory refinement or reacquisition after disturbances in the transfer phase of a mission and during near-Earth tracking. In particular, this is true for range-rate data good to 3 mm/s, as currently quoted for the DSIF, and optical angles with an rms error of 2 s, which are considered possible with existing star trackers, laser telescopes, and attitude stabilization systems. The value of such optical trajectory refinement for a particular mission must, of course, be judged in the context of overall system studies.

The pervading limitations of all specific calculations presented in Sections 2 to 5 is the speculative nature of degradations caused by biases in the astrodynamical models used for the various orbit determinations. These include errors in the astrodynamic and geophysical constants such as the speed of light, station locations, wandering of the Earth's poles, and vagaries in its rate of rotation. The effect of these uncertainties is to render the highest tracking precisions currently available (such as σ_r = 10m at interplanetary ranges) not fully utilizable. A proper treatment of this type of bias involves a significant extension of the analyses presented here and should probably go hand-in-hand with refined modeling of instrument biases encountered in the tracking, data retrieval, and midcourse correction procedures. This requires a more detailed understanding of specific tracking instruments and their performance as a function of range, illumination conditions, etc. Once this level of refinement is contemplated, one should probably go to a full, three-dimensional simulation of critical mission phases, such as the Mars flyby, especially if an autonomous mode of navigation is to be considered.

REFERENCES

1. Bellcomm Apollo Tracking Analysis Program, BTL-R-66-320-1,2,3 (February 1, 1966).
2. On Systematic Errors in Trajectory Determination Problems, A.J. Claus, proceedings, First IFAC Symposium on Automatic Control in the Peaceful Uses of Space, June 21-24, 1965, p. 339.
3. National Bureau of Standards, "Basic Theorems in Matrix Theory," Applied Mathematics Series, No. 59.
4. R.J. Richard and R.Y. Roth, Earth – Mars Trajectories, 1971, JPL Technical Memorandum No. 33, 100, Volume 4, Part B, June 15, 1965.
5. J.D. Anderson, Determination of the Masses of the Moon and Venus and the Astronomical Unit from Radio Tracking Data of the Mariner II Spacecraft, JPL Technical Report 32-816, July 1, 1967.
6. T.W. Hamilton and N.G. Melbourne, Information Content of a Single Pass of Doppler Data from a Distant Spacecraft, JPL Space Program Summary, No. 37-39, Vol. III, p. 18.
7. P.M. Muller, Polar Motion and DSN Station Locations, JPL Space Program Summary, No. 37-45, Vol. III, pp. 10-14.
8. C.J. Vegos and D.W. Trask, Tracking Station Locations as Determined by Radio Tracking Data – Comparison of Results Obtained from Combined Ranger Block III Missions and from Baker – Nunn Optical Data, JPL Space Program Summary No. 37-43, Vol. III, pp. 3-18.
9. Private communication by A.J. Claus on the accepted practice for the analysis of the Apollo mission.
10. R.E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," J. Basic Engineering, Trans ASME, 82D (1960), pp. 33-45.

APPENDIX I. PRIME POWER

Many resources are being devoted to development of adequat power sources for space projects. Table 1-1 (reprinted from Reference 1) summarizes the availability of future prime power sources.

Prime power of 2 to 15 kW will be required just for the transmitters visualized in spacecraft within the next 10- to 15-year period (1 to 5 kW). Solar cells, isotope Brayton, and/or reactor thermoelectric provide adequate power.

Long-life spacecraft have been dependent almost exclusively upon solar cells for power. The solar cell array on the Mariner 4 is the lightest to be designed and flown (as late as May 1967).[1] Its weight/power ratio of 100 lb/kW was one-half of that for the Mariner 2 solar array. Solar cells will continue to be one of the major prime power sources in the immediate future. A functional model of a 12.5-kW panel assembly will soon be completed.[2] An entire system of four of these panels will have a weight/power ratio of 42 lb/kW. This will meet the early goals set by NASA.[1] The long-range goals are 25 lb/kW with power levels up to megawatts.

A serious limitation in the use of solar cells is that the output power is proportional to the Sun's power. Therefore, the suitability of solar cells is dependent on the mission, and it decreases rapidly beyond 1 or 2 A.U. In addition, the solar cell is subject to radiation damage which ultimately degrades its performance.

Nuclear power is generally thought to be the only reasonable source of power in excess of a few tens of kilowatts for more than a few weeks duration.[3] Because of the scarcity (and thus the high cost) of isotopes the reactor is most likely to be used. Much effort is being devoted to the nuclear reactors [1,2] and considerable progress is being made. The weight/power ratio of the reactor is presently reported to be 300 to 400 lb/kW with a goal of 170 lb/kW possible by using the mercury Rankine cycle conversion system.[2,3]

In the range of power from 2 to 10 kW the isotope Brayton system appears to be promising. However, problems exist in developing the packaging; thus, accurate assessments of the mechanical integrity and life capability have not been made. The dates projected by NASA may be optimistic.

Table 1-1

ESTIMATED TECHNOLOGY READY DATES FOR SYSTEMS OF PRINCIPAL INTEREST[1]

	Probable Power Range (kW)				
System	0.01-2.0	2-10	10-25	25-50	>50
Solar cells	1966-1967	1967-1968	1968-1969	1969-1970	
Isotope thermoelectric	1966-1969				
Isotope Brayton		1970-1972			
Reactor thermoelectric			1969-1971		
Reactor Rankine				1972-1974	1976-1980
Reactor thermionic					1976-1980
Batteries					
Sterilizable			1971-1972		
5 year rechargeable			1972-1974		
Fuel Cells, 1 year life			1973-1975		

NASA RN67-1545
1-18-67

To obtain an idea of the relative costs of prime power sources in terms of weight for given missions, weight/power ratios of 75 lb/kW for solar cell arrays and 300 lb/kW for nuclear reactors (both better than actually obtained at the present) will be used. It is noted that future improvement should be expected and these ratios will both be reduced by approximately one-half. For communication system comparisons the relative ratios are the most important, and these appear to be essentially constant (approximately 4) with time.

REFERENCES

1. Transcript of the Briefing for Industry on NASA Space Power and Electric Propulsion Programs, NASA, April 24, 1967.
2. G. Barna. "Power Systems." Space/Aeronautics. Vol. 48 (July 31, 1967), pp 101–106.
3. G. C. Szego, "Space Power Systems State of the Art," Journal of Spacecraft and Rockets (September-October, 1965), pp 641–659.

APPENDIX 2. LARGE ANTENNA RECEIVING ARRAYS

The maximum gain that can be obtained from antennas is limited by the inability to construct and maintain extremely accurate antenna surfaces for microwave frequencies. In the millimeter region, atmospheric distortion of signals and mechanical errors in antenna construction limit gain. Deep space mission requirements make extremely high-gain antenna systems desirable; therefore, some work has been done to determine the feasibility of using arrays of large antennas.[1]

Antenna arrays seem to offer several advantages over a single large antenna of comparable gain. Some of the more important advantages are as follows:

1. The cost of an array is less than that of a single antenna for extremely large equivalent apertures.
2. The antenna gain of an array is not limited to phase errors introduced by the array elements or by the atmosphere under most conditions.
3. Pointing accuracy requirements for each array antenna depend upon individual antenna beam width.
4. Additional gain can be obtained as needed by adding additional antennas.
5. Maintenance can be performed during operation on an antenna, while still operating with a large number of elements, without significantly affecting performance.
6. An array with a large number of elements can suffer failure of an individual element without large gain reduction.
7. The array offers the flexibility of serving several missions simultaneously because it offers multibeam capability.

The array also has several disadvantages, which are as follows:

1. Phase-lock detectors require phase coherent transmission.
2. The number, location, and size of the antennas limit the minimum obtainable elevation angle (because of one or more antennas blocking the others). Therefore, more ground stations could be required.
3. Relatively complicated electronics are required to insure coherent reception in the presence of atmospheric turbulence.
4. Enough signal energy must be available at each antenna to allow the phase-lock loops to track signal phase.
5. Phasefront variations in the signal must be slow enough to allow phase-tracking loops to follow the fluctuations.

At least one receiving array has been built and operated for some time.[2] Operation of the array seems technically feasible; therefore, some effort will be given to determining the cost of large antenna arrays and comparing it to the cost of a single antenna of equivalent gain. Initial cost estimates will be made for single antennas and arrays for operating frequencies of 8, 16, 35, and 94 GHz. System cost will not be calculated because it is a function of the complexity of the ground station, data processing, and storage equipment. However, the basic receiving array components, such as antenna structure and foundation, servo electronics, and receiver electronics, will be considered to determine the optimum antenna diameter and the number of antennas for use in the receiving arrays. Terminal equipment such as computers and recorders must be used regardless of whether one antenna or an array is chosen; therefore, such equipment will not be considered.

Table 2-1 shows the costs assumed for the components selected. The costs shown are budgetary and may rise if extremely sophisticated receivers are employed. The cost of each item can also be expected to drop by 0.90 to 0.95 each time the quantity is doubled (0.95 is used in this report).

Figures 2-1 to 2-4 show the cost of the receiving array as a function of the equivalent diameter of the array. The most important conclusion that can be drawn from these figures is that for each frequency the cheapest array can be chosen. The number of antennas that give the most economical solution for a particular equivalent area varies as a function of the equivalent area, but the optimum antenna size is roughly independent of the equivalent aperture chosen. The fact that the optimum antenna diameter decreases with frequency reflects the cost of increased surface error requirements at higher frequencies. The selection of microwave antennas in the neighborhood of 100 feet in diameter as being optimum is consistent with at least three other estimates.[1,3] Some estimates have placed the optimum antenna diameter as high as 250 feet.[4,5] It should be noted that the primary difference between the estimates is the cost of electronics and

facilities used with each antenna. The highest estimate of optimum antenna diameter was made by considering extremely expensive electronics used with each element of the array. The results of this report are based on simple electronics without redundancy. If more sophisticated equipment is required, larger antennas should be considered to minimize cost.

Table 2-1

COST IN DOLLARS OF SELECTED COMPONENTS OF AN ARRAY

Component	Master	Slave
Antenna	$6.7 \times 10^5 D^{1.3} e^{D/45}$	$6.7 \times 10^5 D^{1.3} e^{D/45}$
Servo electronics	10^5	2.5×10^4
Receiver	10^5	10^5
Total	$2 \times 10^5 + 6.7 \times 10^5 D^{1.3} e^{D/45}$	$1.25 \times 10^5 + 6.7 \times 10^5 D^{1.3} e^{D/45}$

REFERENCES

1. Stanford Research Institute, Feasibility Analysis of a Deep Space Receiving Terminal Array of Large Equivalent Aperture, Final Report, Contract NAS 1-3075 (May, 1964).
2. J. Eberle, High Gain Antenna Array Facilities at the Ohio State University, Contract AF 30(602) - 2166 (September, 1961).
3. J. Eberle, "An Adaptive Phased, Four-Element Array of Thirty-Foot Reflectors for Passive (Echo) Communications Systems," IEEE Transactions on Antennas and Propagation, Vol. AP-12 (March, 1964), pp 169–176.
4. J. Schnader, "Receiving System Design for the Arraying of Independently Steerable Antennas," IRE Transactions on Space Electronics and Telemetry, Vol. SET 9 (June, 1963), pp 148–153.
5. P. Potter, W. Merrick, and A. Ludwig, Large Antenna Apertures and Arrays for Deep Space Communications, Technical Report No. 32-848, Jet Propulsion Laboratory, Pasadena, California (November, 1965).

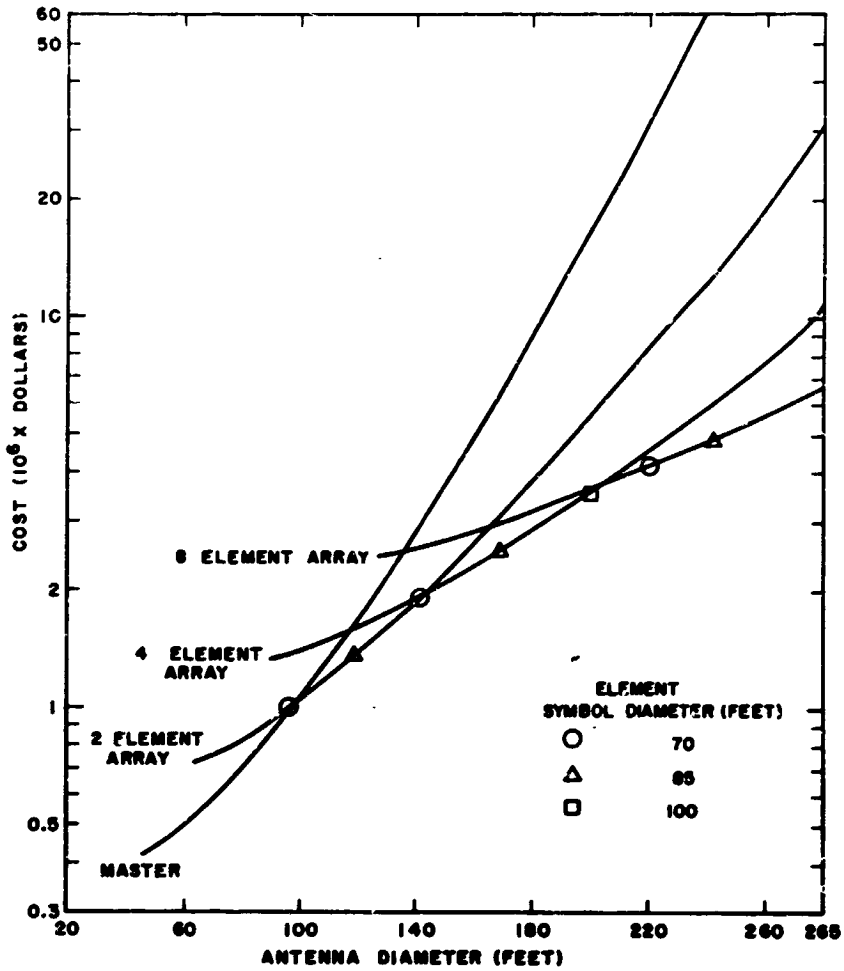


Figure 2-1. Cost of receiving array as a function of equivalent diameter of array (70 to 100 feet)

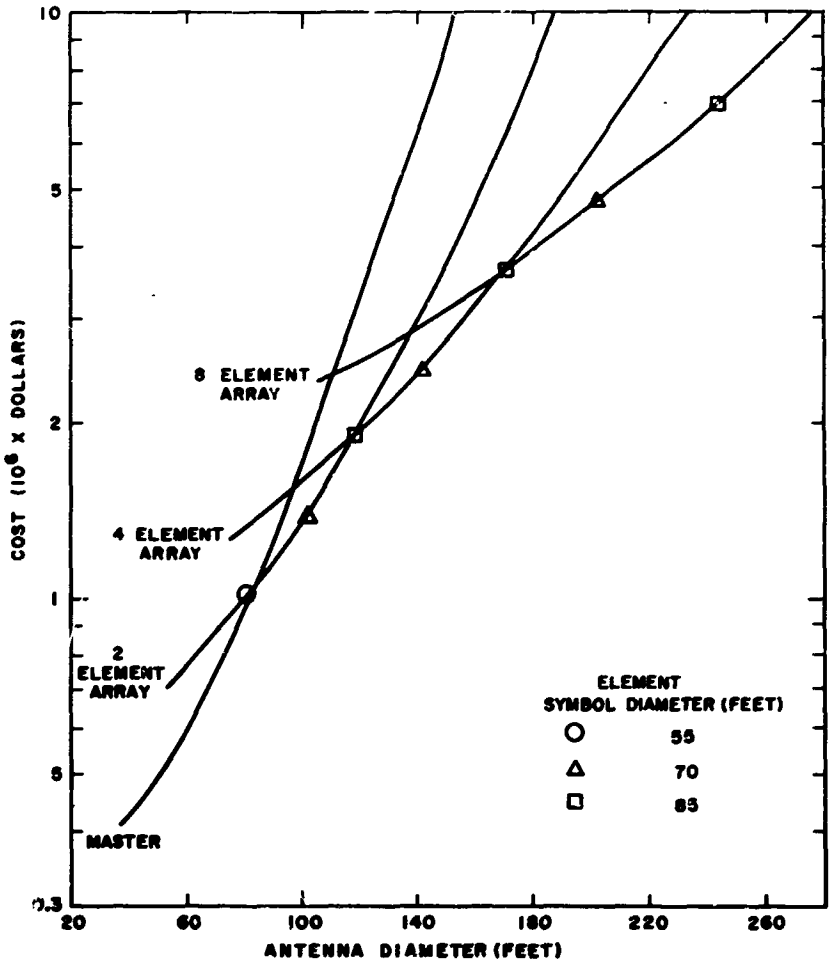


Figure 2-2. Cost of receiving array as a function of equivalent diameter of array (55 to 85 feet)

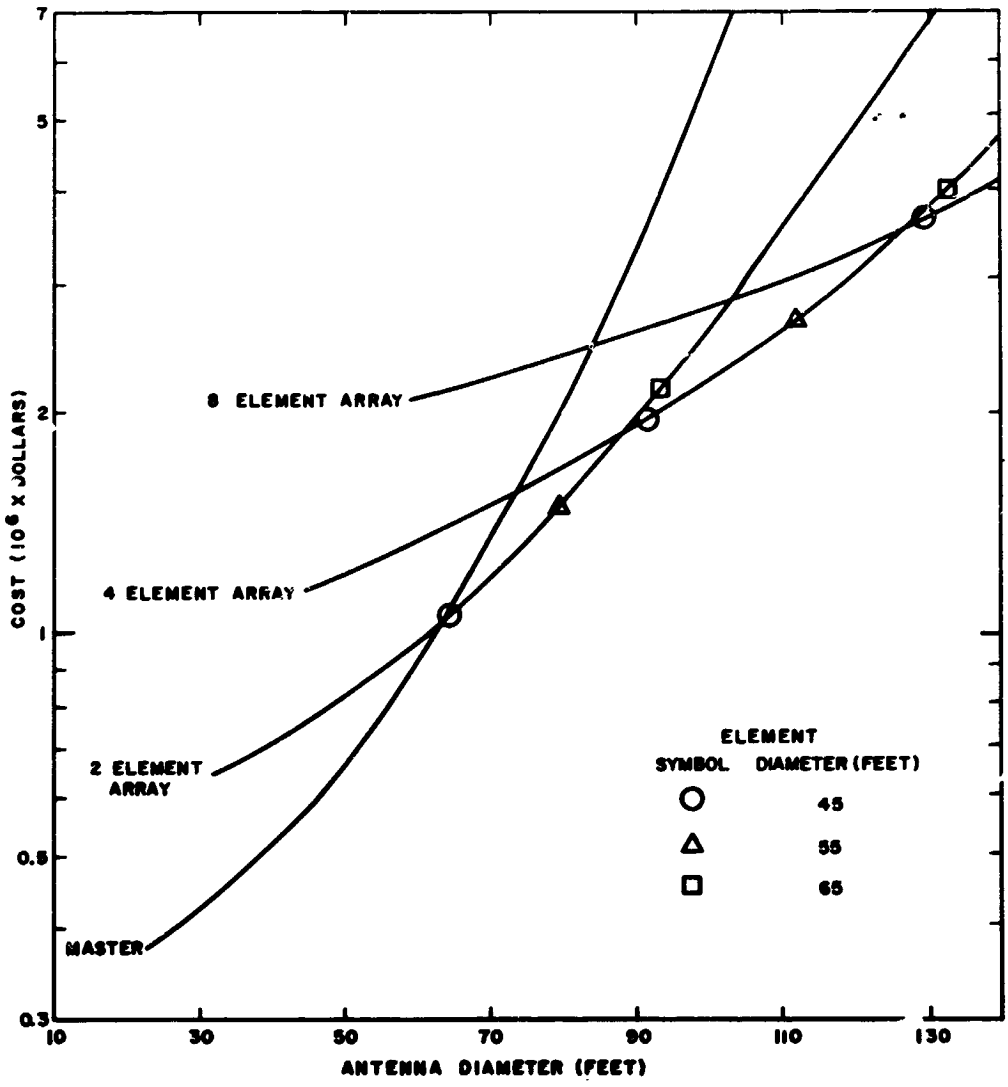


Figure 2-3. Cost of receiving array as a function of equivalent diameter of array (45 to 65 feet)

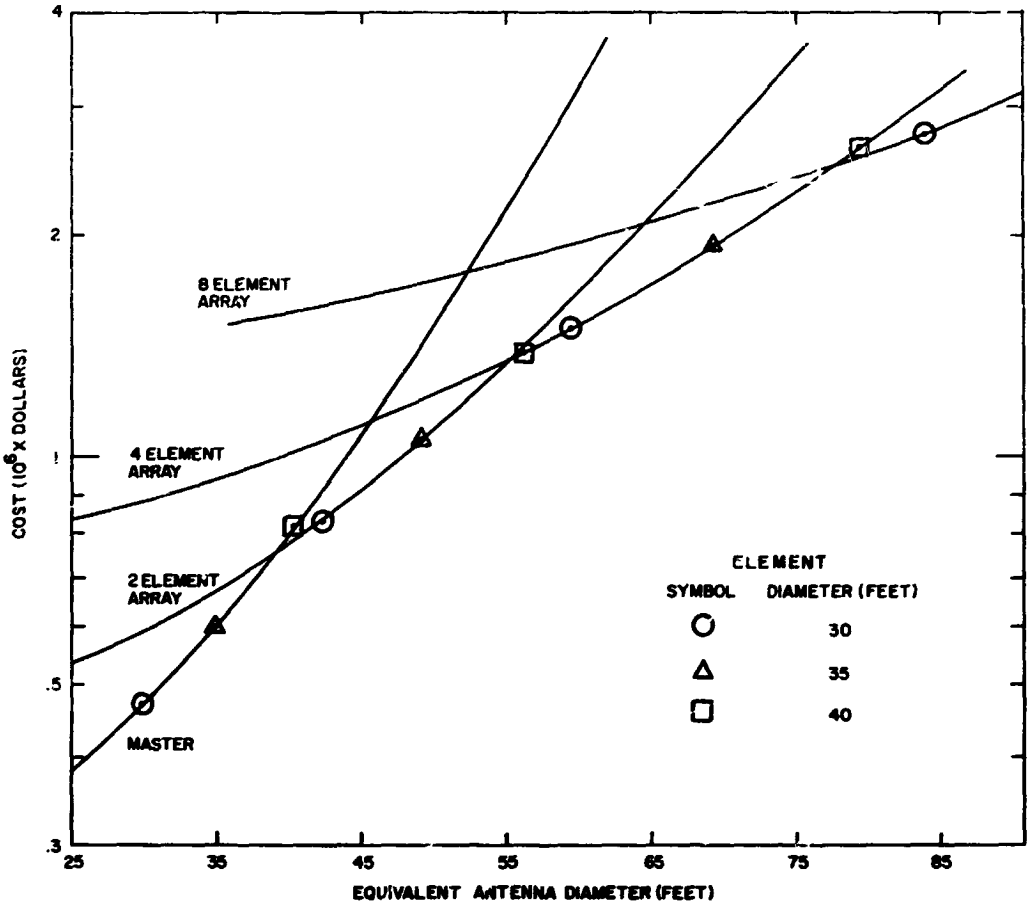


Figure 2-4. Cost of receiving array as a function of equivalent diameter of array (30 to 40 feet)

APPENDIX 3.

GAIN OF ANTENNAS WITH RANDOM SURFACE DEVIATIONS

On-axis gain of antennas with rough reflecting surfaces has been computed as a function of average surface deviation ϵ, correlation distance c, antenna diameter D, and wavelength λ. Gaussian stationary surface deviations, Gaussian correlation functions, and uniform illumination were assumed. It is believed that this represents the first calculation of on-axis antenna gain vs. wavelength when the normalized deviation $(4\pi\epsilon/\lambda)^2$ of the rough surface is larger than 4.

The gain of shallow paraboloid reflector antennas with random surface deviations has been derived by Ruze.[1,2] The derivation was based on a scalar Kirchhoff approximation to the radiation from reflector antennas. The surface deviations were assumed to be Gaussian stationary with Gaussian correlation functions. On these bases an approximate solution for the antenna gain was obtained in terms of an infinite series. The series has been evaluated for relatively small rms surface deviation ϵ in comparison to the wavelength λ, namely $(4\pi\epsilon/\lambda)^2 \leqslant 4$. On-axis gain measurements of large reflector antennas as a function of frequency in general exhibit the characteristics as predicted theoretically by Ruze. Asymptotic limits (as $\lambda \to 0$) for the gain were also given by Ruze[2] based on a similar analysis by Scheffler.[3]

The present work was motivated primarily to determine the gain in the region intermediate between very long and very short wavelengths and to establish a criterion for what surface deviations the asymptotic limit was applicable. Of primary interest was the field distribution, caused by an incident phase wave in the focal plane of a paraboloid reflector antenna in the vicinity of its axis. However, since both the far-field radiation pattern and the field distribution in the focal plane are Fourier transforms of the antenna aperture illumination, the deviations by Ruze are applicable for determining both the far-field and focal-plane distributions.

The series solution for the antenna gain does not seem to be suitable for numerical computations for large values of rms surface deviations. This is because of the large values that some of the terms in the series will assume before the terms begin to decrease. However, it was recognized that, for the on-axis gain, the series is related to an exponential integral. The exponential integral has also an asymptotic series representation, which is suitable for numerical computation for large arguments. On this basis the on-axis gain has been computed as a function of the rms surface deviation to the wavelength ratio and for a range of correlation parameters. The asymptotic limit for the gain is evident from these computations.

The off-axis gain is more difficult to compute, since the asymptotic representations of the series which would facilitate such computations do not seem to be available. However, it is shown that in the asymptotic limit, the gain reduces to that obtained by Scheffler.[3]

In the following sections the gain of antennas with rectangular apertures and Gaussian stationary surface deviations is presented by assuming uniform illumination. A generalization to include certain types of nonuniform illuminations is discussed. The on-axis gain for antennas with circular apertures is also given. It is shown that the on-axis gain for antennas with rectangular and circular apertures can be normalized, such that the normalized gain is the same for both. The asymptotic limit for the off-axis gain is derived. The concluding section summarizes briefly the obtained results.

1. ANTENNA GAIN

The far field gain, $G(\theta,\Phi)$, in the vicinity of the axis of a shallow paraboloid reflector antenna with surface deviations, $z(x,y)$, is, by using the scalar Kirchhoff approximation, given by:[4]

$$G(\theta,\Phi) = \frac{4\pi}{\lambda^2} \frac{\iint_S \iint_S E_a(x,y)E_a^*(x_1,y_1)e^{j\{\beta_x u + \beta_y v + 2k\,[z(x,y)-z(x_1,y_1)]\}} ds\, ds_1}{\iint_S E_a(x,y)E_a^*(x,y)\, ds} \qquad (1)$$

where E_a is the projected electric field on the antenna aperture and s is the aperture area.

$k = \frac{2\pi}{\lambda}$ = free space propagation constant

λ = wavelength

$\beta_x = k \sin\theta \cos\Phi$ (2)

$\beta_y = k \sin\theta \sin\Phi$ (3)

θ and Φ are the spherical coordinates indicated in Figure 3-1.

$$u = x - x_1 \quad (4)$$

$$v = y - y_1 \quad (5)$$

The Kirchhoff approximation is based on the assumption that the surface is locally plane, hence Equation (1) is applicable to surfaces for which the curvatures are small.

Equation (1) can also be used to determine the power distribution in the focal plane of shallow paraboloid reflector antennas, in the vicinity of the focal point, in which case (referring to Figure 3-1)

$$\beta_x = kx_f/f \quad (6)$$

$$\beta_y = ky_f/f \quad (7)$$

where x_f and y_f are the coordinates in the focal plane and f is the focal length.

If z(x,y) is a Gaussian stationary random variable with zero mean it has been shown[1,5] that, by performing the statistical averaging, the expectation value for the gain, $\overline{G(\theta,\Phi)}$, is:

$$\overline{G(\theta,\Phi)} = \frac{4\pi}{\lambda^2} e^{-\delta^2} \frac{\iint_S \iint_S E_a(x,y)E_a^*(x_1,y_1)e^{j(\beta_x u + \beta_y v)} e^{\delta^2 r(u,v)}\, ds\, ds_1}{\iint_S E_a(x,y)E_a^*(x,y)\, ds} \quad (8)$$

where

$$\delta = \frac{4\pi}{\lambda}\epsilon$$

ϵ = rms surface deviation

$\delta^2 r(u,v)$ = correlation function. (9)

To evaluate Equation (8), four integrations have to be performed. It is shown in Attachment A that for antennas with rectangular apertures two integrations can be readily eliminated for certain types of illuminations as, for example, truncated cosine illuminations.

In particular for uniform illuminations, $E_a(x,y) = 1$, and for a Gaussian correlation function, r, defined as

$$r(u,v) = e^{-\left(\frac{u^2+v^2}{c^2}\right)} \quad (10)$$

where c is the correlation length, it is shown in Attachment A that the expectation value of the gain is:

$$\overline{G(\theta,\Phi)} = e^{-\delta^2} G_o(\theta,\Phi) + \left(\frac{2\pi c}{\lambda}\right)^2 e^{-\delta^2} \sum_{n=1}^{\infty} \frac{\delta^{2n}}{n!n} \left[e^{-\beta^2 c^2/4n} - \Delta_n\right] \quad (11)$$

where $G_o(\theta,\Phi)$ is the antenna gain in the absence of surface deviations.

For an antenna with a rectangular aperture

$$G_o(\theta,\Phi) = \frac{4\pi A}{\lambda^2}\left(\frac{\sin\beta_x a}{\beta_x a}\,\frac{\sin\beta_y b}{\beta_y b}\right)^2 \quad (12)$$

where A = 4ab is the aperture area.

$$\beta = \frac{2\pi}{\lambda}\sin\theta \quad (13)$$

and

$$\Delta_n < \frac{c}{2\sqrt{\pi n}}\left[\frac{1}{a} + \frac{1}{b}\right] \quad (14)$$

Equation (11) is in agreement with the gain derived by Ruze for antennas with circular apertures except for the term Δ_n. This term is small, since the correlation distance c is small in comparison to the linear dimensions of the antenna.

For antennas with circular apertures the evaluation of Equation (8) is in general more difficult. However, the on-axis gain $\overline{G(0,0)}$ has been derived in Attachment B for uniform illumination. The gain has the same functional form as Equation (11) with $\beta = 0$ and

$$\Delta_n \approx \frac{2c}{D\sqrt{\pi n}} \quad (15)$$

where D is the antenna aperture diameter.

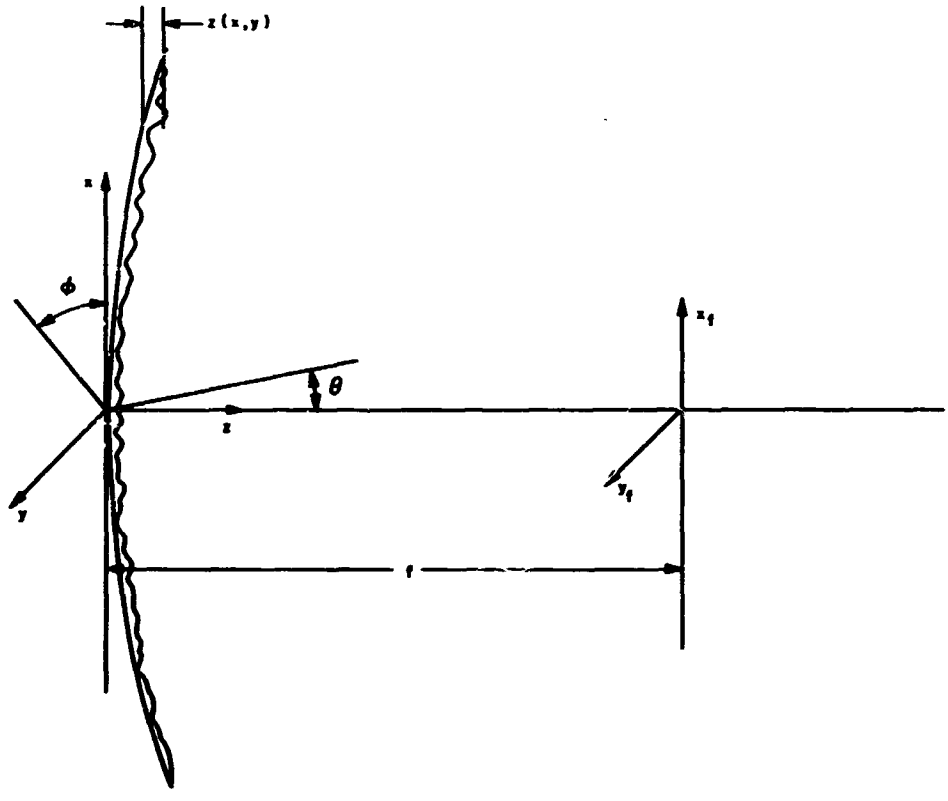


Figure 3-1. Antenna coordinates

2. ON-AXIS GAIN

Equation (11) can be readily computed for small values of δ^2. For large values of δ^2 the terms $(\delta^2)^n / n!$ will assume very large values, which would exceed the accuracy of present computers before the terms decrease. This series is therefore not suitable for direct computation of if δ^2 is large. However, the gain on-axis can be readily computed by noticing that the series in Equation (8) for $\rho = 0$ is related to an exponential integral, which also has an asymptotic representation.

The exponential integral, E_i, can be written,[6]

$$E_i(x) = \gamma + \ln x + \sum_{n=1}^{\infty} \frac{x^n}{n!n} \tag{16}$$

where γ is Euler's constant. The asymptotic series ($x \to \infty$) for $E_i(x)$ is

$$E_i(x) = \frac{e^x}{x} \sum_{n=0}^{N-1} \left[\frac{n!}{x^n} + 0 \left(\frac{1}{x^N} \right) \right] \tag{17}$$

Though the asymptotic series diverges for all finite values of x it can be used to evaluate $E_i(x)$ for large x by using up to N terms[7], where N is an integer nearest to the value of x.

In terms of the exponential integral, the on-axis gain for both rectangular and circular aperture antennas is

$$\overline{G(0,0)} = \left(\frac{D_o}{4\epsilon}\right)^2 \left\{ \delta^2 e^{-\delta^2} + \left(\frac{2c}{D_o}\right)^2 \delta^2 e^{-\delta^2} \left[E_i(\delta^2) - \ln \delta^2 - \gamma \right] \right\} \tag{18}$$

where D_o is related to the antenna area, A, by

$$A = \frac{\pi D_o^2}{4} \tag{19}$$

One parameter in Equation (18) is readily eliminated by defining a normalized on-axis gain, $\overline{g(0,0)}$, by

$$\overline{g(0,0)} = \frac{\overline{G(0,0)}}{(D_o/4\epsilon)^2} = \delta^2 e^{-\delta^2} + \left(\frac{2c}{D_o}\right)^2 \delta^2 e^{-\delta^2} \left[E_i(\delta^2) - \ln \delta^2 - \gamma \right] \tag{20}$$

The normalized gain depends only on two parameters δ^2 and $(c/D_o)^2$.

Equation (20) has been computed by using a SHARE program for the computation of the exponential integral developed by D. S. Villars. This program computes $E_i(x)$ with at least 4 decimal accuracy.

Computations have been performed for $10^{-4} \leq \delta^2 \leq 80$ and for $10^{-3} \leq c/D_o \leq 0.1$. The computed normalized gain is shown in Figure 3-2.

The computations show the normalized antenna gain has three distinct regions which are characterized by the normalized rms surface deviation to wavelength ratio δ.

In the region $0 \leq \delta^2 \leq 1$ the normalized antenna gain is nearly independent of the correlation length, and increases almost linearly with δ^2. In the region $1 \leq \delta^2 \leq 20$ the gain is dependent on both δ^2 and c. In the region $\delta^2 > 20$ the gain is almost independent of δ^2 and is a function of c/D_o only. This region is the asymptotic region. For the range of parameters used in the computation, the gain in the asymptotic region, for a given c/D_o ratio, deviates by less than 5 percent from the asymptotic value.

The curves shown in Figure 3-2 seem to confirm the general characteristics of the measured gain as a function of frequency of large reflector antennas presented by Ruze,[2] which is included here as Figure 3-3.

The presented measurements extend only slightly from the first into the second region but not sufficiently far to show the asymptotic region. A detailed comparison of the measured and computed gain cannot be made since uniform illumination has been assumed in the computation.

3. ASYMPTOTIC VALUES FOR THE GAIN OFF AXIS

The computation of the off-axis gain directly from Equation (11) can only be performed for relatively small values of δ^2. An alternate representation for the gain is obtained by expanding the exponential in the second term of Equation (11) in a power series. By using this expansion and neglecting Δ_n, Equation (11) can be rewritten as follows:

$$\overline{G(\theta,\Phi)} = e^{-\delta^2} G_o(\theta,\Phi) + \left(\frac{c}{2\epsilon}\right)^2 \sum_{m=0}^{\infty} (-1)^m \frac{A_m}{m!} \left(\frac{\beta_c}{2\delta}\right)^{2m} \tag{21}$$

where

$$A_m = e^{-\delta^2} \sum_{n=1}^{\infty} \frac{(\delta^2)^{n+m+1}}{n!\, n^{m+1}} \tag{22}$$

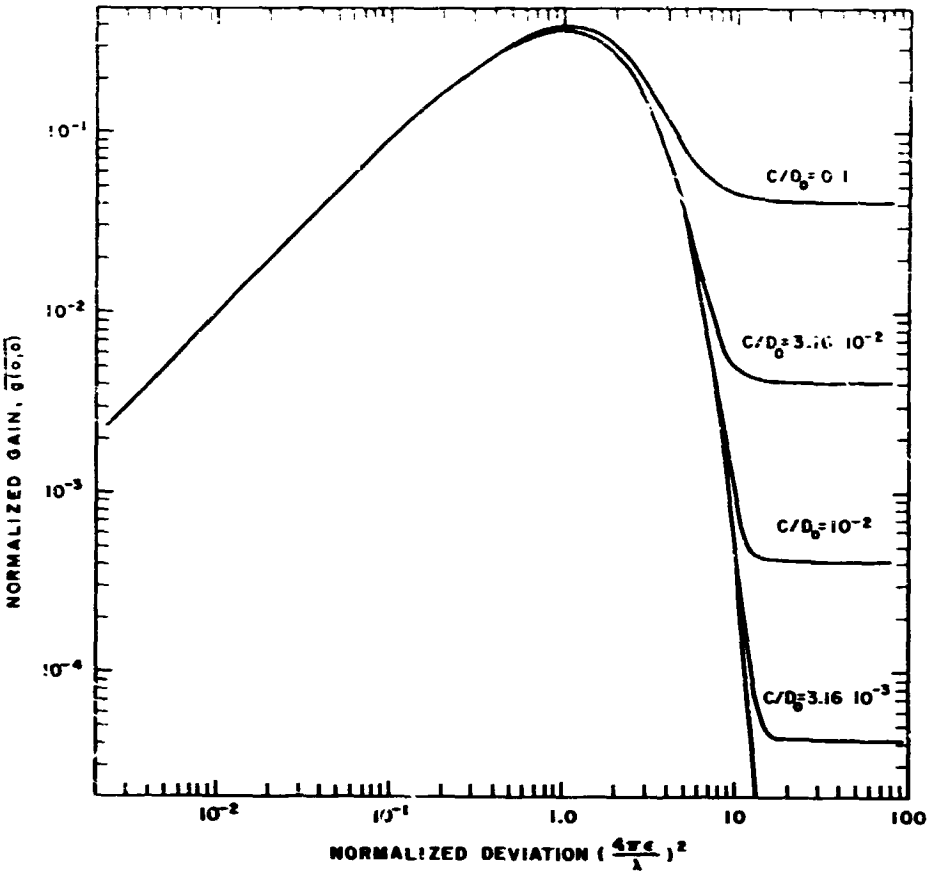


Figure 3-2. Normalized antenna gain

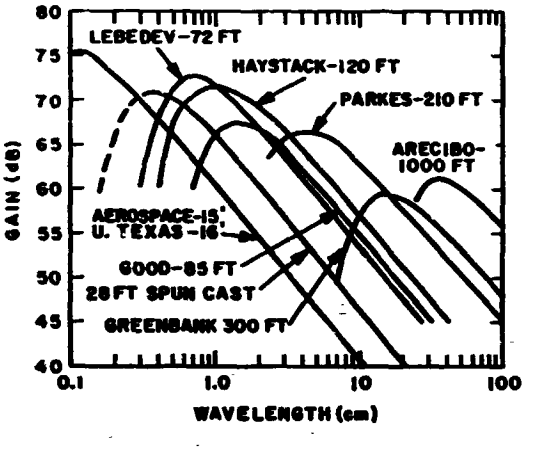


Figure 3-3. Gain of large paraboloids (reproduced from a paper by Ruze[2])

Except for the special case m = 0 treated above, asymptotic series representations for A_m do not seem to be available. The first term of the asymptotic series has been obtained in Attachment C, where it is shown that for $\delta^2 \to \infty$

$$A_m = 1 + 0\left(\frac{1}{\delta^2}\right) \tag{23}$$

The off-axis asymptotic gain will be designated by $\overline{G(\theta,\Phi)_\infty}$ and is given by:

$$\overline{G(\theta,\Phi)_\infty} = \left(\frac{c}{2\epsilon}\right)^2 e^{-(c/4\epsilon \sin\theta)^2} \tag{24}$$

and the corresponding normalized gain

$$\overline{g(\theta,\Phi)_\infty} = \left(\frac{2c}{D_o}\right)^2 e^{-(c/4\epsilon \sin\theta)} \tag{25}$$

The asymptotic value for the gain, Equation (24), is in agreement with the gain obtained by Scheffler,[3] based on the following approximation to the Gaussian correlation function.

$$\exp\left[-\delta^2\left(1-e^{-\frac{u^2+v^2}{c^2}}\right)\right] = \begin{cases} \left(1-e^{-\delta^2}\right)e^{(u^2+v^2)/c^2} + e^{-\delta^2} & \delta^2 \le 1 \\ \left(1-e^{-\delta^2}\right)e^{-\delta^2/c^2(u^2+v^2)} + e^{-\delta^2} & \delta^2 \le 1 \end{cases} \tag{26}$$

Equation (24) is independent of frequency but is strongly dependent on the ratio ϵ/c. This ratio has been interpreted as average surface slope.[2]

A comparison can be made between the previously computed on-axis gain and that obtainable by using the approximation in Equation (26). With the latter the first term in Equation (20) is the same, and the second term is approximated by

$$\delta^2 e^{-\delta^2}\left[E_1(\delta^2) - \ln\delta^2 - \gamma\right] \approx \begin{cases} \delta^2\left(1-e^{-\delta^2}\right) & 0 \le \delta^2 \le 1 \\ 1 - e^{-\delta^2} & \delta^2 \ge 1 \end{cases} \tag{27}$$

The R.H.S. and the L.H.S. of Equation (27) are shown in Figure 3-4 as a function of δ^2. The maximum deviation is 23.4 percent at $\delta^2 = 4$.

The asymptotic region for the off-axis gain has not been determined precisely, however it is reasonable to assume that this region will correspond to the asymptotic region for the on-axis gain.

To obtain an estimate for the off-axis gain, the gain of a uniformly illuminated circular aperture antenna without surface deviations ($\delta^2 = 0$) is compared with gain of such an antenna with $\delta^2 = 25$.

For $\delta^2 = 0$, the gain can be written

$$\frac{\overline{G_o(\theta,\phi)}}{\left(\frac{\pi D}{\lambda}\right)^2} = \left(\frac{2J_1(W)}{W}\right)^2 \tag{28}$$

J_1 is a Bessel function of order one.

$$W = \frac{\pi D}{\lambda}\sin\theta \tag{29}$$

For $\delta^2 = 25$, by using the asymptotic values for the gain

$$\frac{G(\theta,\phi)}{\left(\frac{\pi D}{\lambda}\right)^2} \approx \left(\frac{2c}{D\delta}\right)^2 e^{-(c/D\, W/\delta)^2} \tag{30}$$

Figure 3-5 shows a graph of Equation (28) as a function of the normalized radius W. In the same figure is also shown a graph of Equation (30) for c/D = 31.6 and c/D = 100. The increase of the beamwidth with δ^2 and the strong dependence of the beamwidth on the ratio of c/D is apparent from this figure.

4. CONCLUSIONS

The on-axis gain of antennas with Gaussian stationary random surface deviations and Gaussian correlation functions has been determined for antennas with rectangular and circular apertures by assuming uniform illumination. For both types a normalized expression for the gain was derived which depends only on the normalized rms surface deviation to wavelength ratio, δ, and the ratio of the correlation length c to a defined linear antenna dimension D_o. For circular antennas, D_o is the diameter.

The antenna gain as a function of δ exhibits three distinct regions: (1) $0 \le \delta^2 \le 1$, (2) $1 \le \delta^2 \le 20$, and (3) $\delta^2 > 20$. The last region is designated as the asymptotic region. In this region the gain is nearly independent of wavelength.

The computed gain exhibits in general the characteristics of the measured gain as a function of frequency of large reflector antennas reported in the literature. These measurements extend only into the

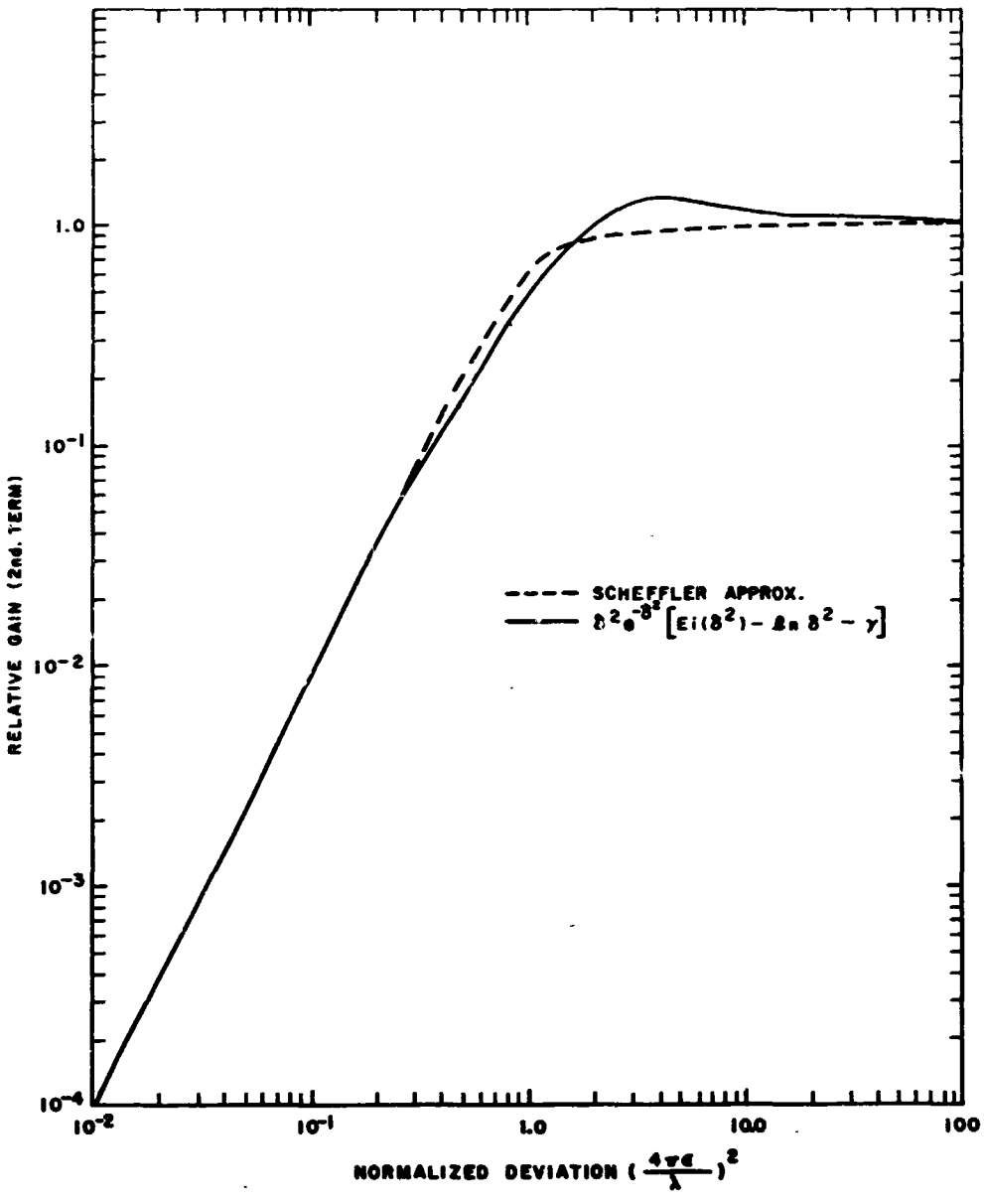


Figure 3-4. Comparison of the second terms for the on-off gain

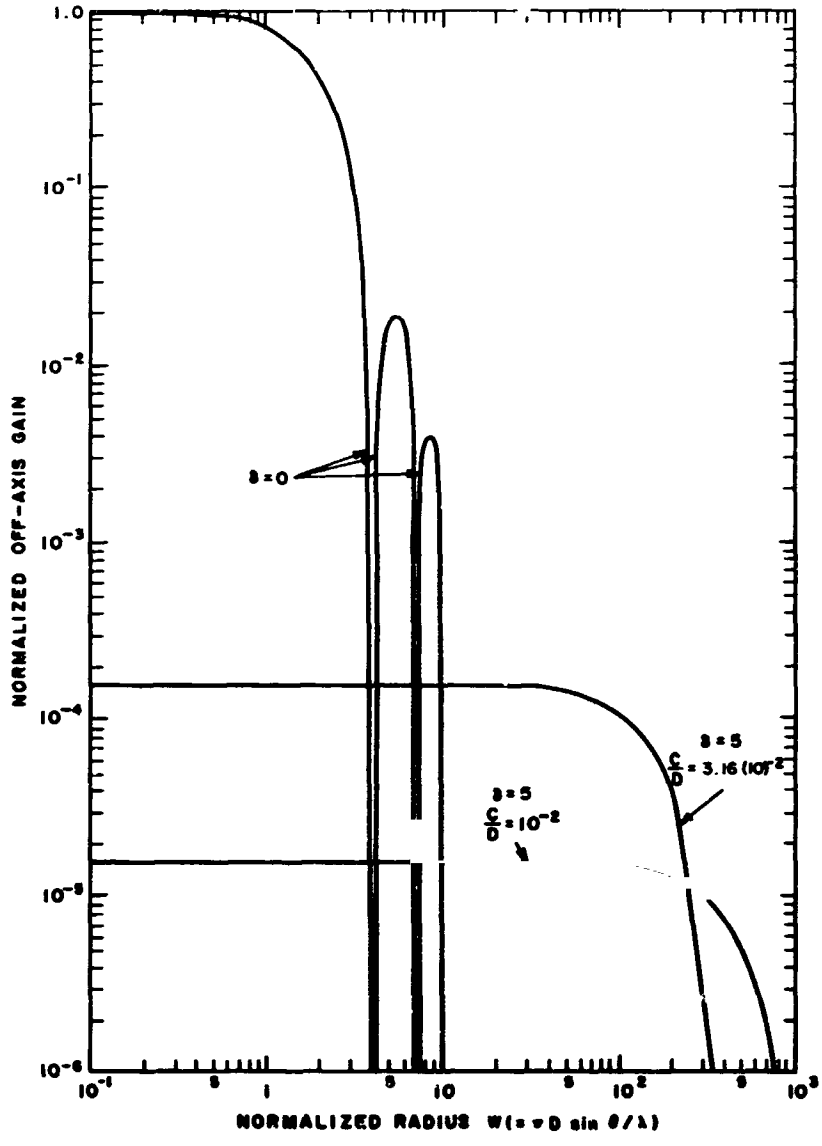


Figure 3-5. Off-axis antenna gain

second region and therefore not far enough to show the third (asymptotic) region.

The off-axis gain can be readily determined in the first and third regions. It is shown that, in the third region, the expression for the gain reduces to that previously obtained by Scheffler. In the second region the off-axis gain computation is more difficult, since the desired representation of the series which would facilitate the computations do not seem to be available.

REFERENCES

1. J. Ruze, "The Effect of Aperture Errors on the Antenna Radiation Pattern," Suppl. al Nuovo Cimento, Vol. 9, No. 3 (1952), pp 364–380.
2. J. Ruze, "Antenna Tolerance Theory – A Review," Proc. IEEE, Vol. 54, No. 4 (April 1966), pp 633–640.
3. H. Scheffler, "Uber die Genauigkeitsforderungen bei der Herstellung Optischer Flachen fur Astronomische Teleskope," Z. Astrophys. (Germany), Vol. 55 (1962), pp 1–20.
4. S. Silver, Microwave Antenna Theory and Design (New York, McGraw-Hill, 1949).
5. W. C. Hoffman, "Scattering of Electromagnetic Waves from a Random Surface," Quarterly of Appl. Math., Vol. 13, No. 3 (1955), pp 291–304.
6. A. Erdelyi et al, Higher Transcendental Functions, Vol. 2 (New York, McGraw-Hill, 1953), pp 143–144.
7. P. H. Morse and H. Feshbach, Methods of Theoretical Physics, Part I (New York, McGraw-Hill, 1953), pp 434–443.
8. Y. L. Luke, Integrals of Bessels Functions (New York, McGraw-Hill, 1962), pp 271–283.
9. G. N. Watson, A Treatise on the Theory of Bessel Functions, 2nd ed. (Cambridge, The University Press, 1962), pp 393–395.

Attachment A. GAIN OF ANTENNAS WITH RECTANGULAR APERTURES

To evaluate Equation (8) for antennas with rectangular apertures, consider the following integral

$$\bar{I} = e^{-\delta^2} \int_{-a}^{a} \int_{-a}^{a} \int_{-b}^{b} \int_{-b}^{b} E_a(x,y) E_a(x_1,y_1) e^{\delta^2 r(u,v)} e^{j(\beta_x u + \beta_y v)} dx\, dx_1\, dx\, dy_1 \qquad (A\text{-}1)$$

with

$$u = x - x_1 \qquad (A\text{-}2)$$

$$v = y - y_1 \qquad (A\text{-}3)$$

Since Equation (A-1) contains the correlation function in terms of u and v, it is preferable to introduce the u,v coordinate system.

In the x,x_1 coordinate system the integrations are over the square region shown in Figure 3-6. In the y,y_1 system the region is similar. With the coordinate transformation Equation (A-2), the transformed region in the x_1,u coordinate system is also shown in Figure 3-6.

In the x_1,u plane integration with respect to x_1 is readily performed for certain types of illumination functions.* In particular, let

$$E_a(x,y) = E_{ax}(x) E_{ay}(y) \qquad (A\text{-}4)$$

and

$$E_a(x_1+u) = \sum_{n=1}^{N} f_n(x_1) g_n(u) \qquad (A\text{-}5)$$

with a similar equation for $E_a(y_1+v)$. An example of such an illumination is a truncated cosine illumination, where Equation (A-5) will consist of two terms.

It is sufficient to consider the following integral

$$I_1 = \int_{-a}^{a} \int_{-a-x_1}^{a-x_1} g(x_1) f(u,v)\, du\, dx_1 \qquad (A\text{-}6)$$

*A similar method has been used by Hoffman in his treatment of scattering of electromagnetic waves from a random surface[3].

Referring to Figure 3-6, Equation (A-6) can be written as

$$\bar{I}_1 = \int_{0}^{2a} \int_{-a}^{a-u} g(x_1) f(u,v)\, dx\, du + \int_{-2a}^{0} \int_{-a-u}^{a} g(x_1) f(u,v)\, dx_1\, du \qquad (A\text{-}7)$$

let $G(x_1) = \int g(x_1)\, dx$

then

$$\bar{I}_1 = \int_{0}^{2a} \Big\{ [G(a-u) - G(-a)]\, f(u,v) + [G(a) - G(-a+u)]\, f(-u,v) \Big\} du \qquad (A\text{-}8)$$

Using Equation (A-8) and assuming uniform illumination, $E_a = 1$, two integrations are readily eliminated and Equation (A-1) reduces to

$$\bar{I} = 4e^{-\delta^2} \int_{0}^{2a} \int_{0}^{2b} (2a-u)(2b-v)\, e^{\delta^2 r(u,v)} \cos\beta_x u \cos\beta_y v\, du\, dv \qquad (A\text{-}9)$$

By expanding the exponential function in a power series, Equation (A-9) can be divided into two parts corresponding to the coherent and incoherent parts of gain, as follows

$$\bar{I} = I_c + I_{inc} \qquad (A\text{-}10)$$

and

$$I_c = (\bar{I})^2 = A^2 e^{-\delta^2} \left(\frac{\sin\beta_x a}{\beta_x a} \frac{\sin\beta_y b}{\beta_y b} \right)^2 \qquad (A\text{-}11)$$

and

$$I_{inc} = 4A \sum_{n=1}^{\infty} \int_{0}^{2a} \int_{0}^{2b} \frac{(\delta^2 r(u,v))^n}{n!} \cos\beta_x u \cos\beta_y v\, du\, dv - A I_{inc} \qquad (A\text{-}12)$$

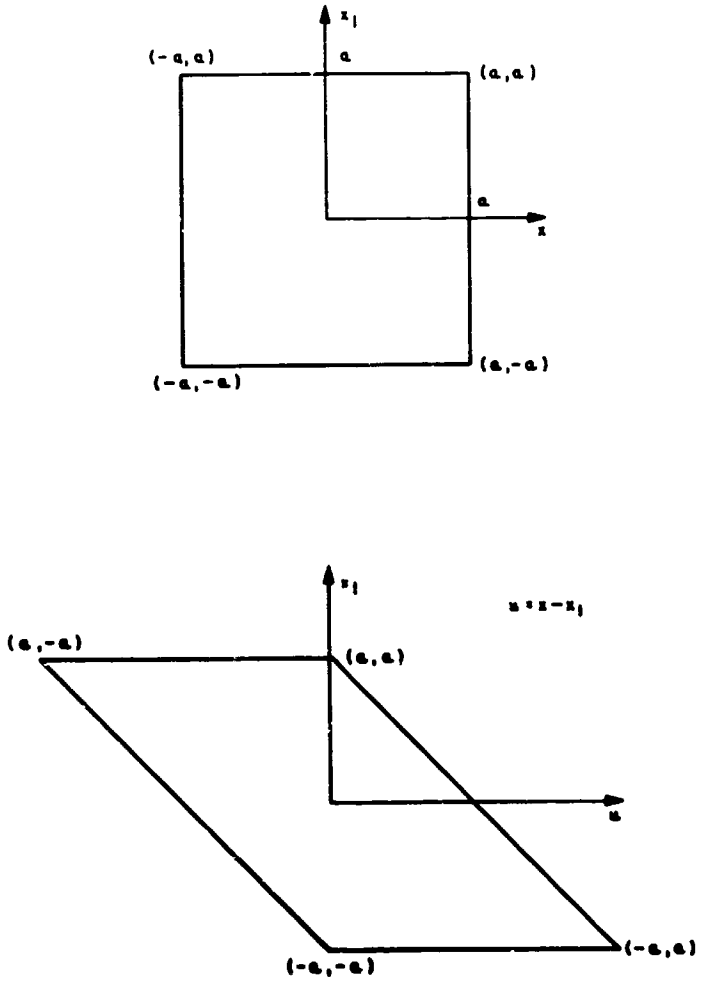


Figure 3-6. Coordinate transformation

where

$$\Delta\bar{I}_{unc} = 4e^{-\delta^2} \sum_{n=1}^{\infty} \int_0^{2a} \int_0^{2b} [2(bu+av)-uv] \frac{[\delta^2 r(u,v)]^n}{n!} \cos\beta_x u \cos\beta_y v \, du dv \qquad (A\text{-}13)$$

and $A = 4ab$ is the aperture area.

It is noted that the coherent part of the gain is the same as the antenna gain in the absence of surface deviations but multiplied by $e^{-\delta^2}$. This follows from Equation (A-1) by expanding the exponential function which contains the correlation function in a power series.

To obtain an estimate for I_{unc}, Equation (A-12) is evaluated for $\beta_x = \beta_y = 0$, corresponding to the on-axis gain, and for a Gaussian correlation function

$$r(u,v) = e^{-(u^2+v^2)/c^2} \qquad (A\text{-}14)$$

where c is the correlation distance.

On-axis

$$I_{unc}(0,0) = \pi A c^2 \sum_{n=1}^{\infty} \frac{(\delta^2)^n}{n!n} \left(1 - \frac{c}{2\sqrt{\pi n}}\left(\frac{1}{a}+\frac{1}{b}\right) + \frac{c^2}{\pi A n}\right) \qquad (A\text{-}15)$$

In Equation (A-15) terms of order $e^{-n(2a/c)^2}$ and $e^{-n(2b/c)^2}$ were neglected.

By extending the limits of integrations in Equation (A-12) to ∞, the integration of the first part of this equation can be performed and gives Equation (11) of the text.

Attachment B. ON-AXIS GAIN FOR CIRCULAR APERTURE ANTENNAS

For circular aperture antennas the on-axis gain for uniform illumination and a Gaussian correlation function is obtained from Equation (8) by expanding the exponential function and performing the integrations for the n=0 terms, and the integration with respect to the azimuthal coordinates for the remaining terms, resulting in

$$\overline{G(O,O)} = \left(\frac{\pi D}{\lambda}\right)^2 e^{-\delta^2} + \left(\frac{8\pi}{\lambda D}\right)^2 \sum_{n=1}^{\infty} \frac{(\delta^2)^n}{n!} I_{cn} \qquad \text{(B-1)}$$

where D is aperture diameter, and

$$I_{cn} = \int_0^{D/2} \int_0^{D/2} e^{-n(\rho^2+\rho_1)/c^2} I_o\left(\frac{2n\rho\rho_1}{c^2}\right) \rho \, d\rho\rho_1 \, d\rho_1 \qquad \text{(B-2)}$$

I_o - Modified Bessel function of order zero.

The two integrations in Equation (B-2) will be performed with the aid of the $Q(y,a_n)$ function defined by[8]

$$Q(y,a_n) = \int_{a_n}^{\infty} e^{-(x^2+y^2)/2} I_o(xy) x \, dx \qquad \text{(B-3)}$$

Let

$$x = \sqrt{2n}\,\rho/c \qquad \text{(B-4)}$$

$$y = \sqrt{2n}\,\rho_1/c \qquad \text{(B-5)}$$

$$a_n = \frac{D}{c}\sqrt{\frac{n}{2}} \qquad \text{(B-6)}$$

With Equation (B-3) - Equation (B-6), Equation (B-2) can be written

$$I_{cn} = \left(\frac{c^2}{2n}\right)^2 \int_0^{a_n} \left[1 - Q(y,a_n)\right] y \, dy \qquad \text{(B-7)}$$

Integrating by parts results in

$$I_{cn} = \left(\frac{c^2}{2n}\right)\left\{\frac{a_n^2}{2}\left[1 - Q(a_n,a_n)\right] + \int_0^{a_n} \frac{y^2}{2} \frac{\partial Q}{\partial y} dy\right\} \qquad \text{(B-8)}$$

The derivative in Equation (B-8) can be expressed as

$$\frac{\partial Q}{\partial y} = a_n e^{-(a_n^2+y^2)/2} I_1(a_n y) \qquad \text{(B-9)}$$

Equation (B-9) is readily derived from Equation (B-3) and the following integral[9]

$$\int_0^{\infty} e^{-t^2/2} J_v(xt) J_v(yt) t \, dt = e^{-x^2+y^2/2} I_v(xy) \qquad \text{(B-10)}$$

where J_v is a Bessel function of order v. Substituting Equation (B-9) into Equation (B-8) and integrating by parts results in

$$I_{cn} = \left(\frac{c^2}{2n}\right)^2 \left\{\frac{a_n^2}{2}\left[1 - Q(a_n,a_n) - e^{-a_n^2} I_1\left(a_n^2\right)\right] + \frac{a_n}{2} \int_0^{a_n} e^{-\frac{1}{2}(a_n^2+y^2)} \frac{d}{dy}\left[y I_1(a_n y)\right] dy\right\} \qquad \text{(B-11)}$$

Using the relation

$$\frac{d}{dy} y I_1(a_n y) = a_n y I_o(a_n y) \qquad \text{(B-12)}$$

and the definition of the Q function, Equation (B-3), results in

$$I_{cn} = \left(\frac{c}{2n}\right)\left\{a_n^2\left[1 - Q(a_n,a_n)\right] - \frac{a_n^2}{2} e^{-a_n^2} I_1\left(a_n^2\right)\right\} \qquad \text{(B-13)}$$

To evaluate $Q(a_n,a_n)$ use is made from the following relation readily derived from Equations (B-3) and (B-10)

$$Q(\alpha,\beta) + Q(\beta,\alpha) = 2 + \int_0^{\infty} e^{-\frac{1}{2}t^2} \left\{\frac{d}{dt}\left[J_o(\alpha t) J_o(\beta t)\right]\right\} dt \qquad \text{(B-14)}$$

Integrating Equation (B-14) by parts and using Equation (B-10) gives

$$Q(y,x) + Q(x,y) = 1 + e^{-x^2+y^2/2} I_o(xy) \qquad \text{(B-15)}$$

With Equations (B-13) and (B-6), Equation (B-2) is given by

$$I_{cn} = \left(\frac{cD}{4}\right)^2 \frac{1}{n}\left\{1 - e^{-n/2(D/c)^2}\left[I_o\left(\frac{n}{2}\frac{D^2}{c^2}\right) + I_1\left(\frac{n}{2}\frac{D^2}{c^2}\right)\right]\right\} \qquad \text{(B-16)}$$

The gain on axis (B-1) can therefore be written by using Equation (B-11) as:

$$\overline{G(0,0)} = \left(\frac{D}{4\epsilon}\right)^2 \left[\delta^2 e^{-\delta^2} + \left(\frac{2c}{D}\right)^2 \delta^2 e^{-\delta^2} \sum_{n=1}^{\infty} \frac{\delta^{2n}}{n!n} \left[1 - \Delta_n\right]\right] \qquad \text{(B-17)}$$

$$\text{with } \Delta_n = e^{-\frac{n}{2}\left(\frac{D}{c}\right)^2} \left[I_0\left(\frac{n}{2}\frac{D^2}{c^2}\right) + I_1\left(\frac{n}{2}\frac{D^2}{c^2}\right)\right] \qquad \text{(B-18)}$$

For $n/2(D/c)^2 >> 1$, when the modified Bessel functions can be approximated by the first terms of the asymptotic series, Δ_n is then given by Equation (15) of the text.

Attachment C. ASYMPTOTIC VALUES FOR THE GAIN

Consider the series, A_m, given by Equation (22) in the text, which enters in the evaluation of the off-axis gain. This series can be written as:

$$A_m = e^{-\delta^2} \sum_{n=1}^{\infty} \frac{(\delta^2)^{n+m+1}(n+1)(n+2)(n+3)\ldots(n+m+1)}{n!n^{m+1}(n+1)(n+2)(n+3)\ldots(n+m+1)} \quad \text{(C-1)}$$

Performing the indicated multiplications in the numerator of Equation (C-1), A_m is rewritten as follows

$$A_m = e^{-\delta^2}\left[\sum_{n=1}^{\infty} \frac{(\delta^2)^{n+m+1}}{(n+m+1)!} + \sum_{\ell=1}^{m+1} \sum_{n=1}^{\infty} a_\ell \frac{(\delta^2)^{n+m+1}}{(n+m+1)!n^\ell} \right] \quad \text{(C-2)}$$

where a_ℓ are constants.

The first series in Equation (C-2) is summable since

$$\sum_{n=1}^{\infty} \frac{(\delta^2)^{n+m+1}}{(n+m+1)!} = e^{\delta^2} - \sum_{k=0}^{m+1} \frac{(\delta^2)^k}{k!} \quad \text{(C-3)}$$

The second series in Equation (C-2) is estimated for $\delta^2 \to \infty$ by proceeding in a similar manner as in Equation (C-1); hence,

$$\sum_{n=1}^{\infty} \frac{(\delta^2)^{n+m+1}}{(n+m+1)!n^\ell} = \frac{e^{\delta^2}}{(\delta^2)^\ell}\left[1 + O\left(\frac{1}{\delta^2}\right)\right] \quad \text{(C-4)}$$

From Equations (C-3) and (C-4) the asymptotic value for A_m is obtained and is given by Equation (23) of the text.

APPENDIX 4. STABILITY CONDITION OF THE BEAM-POINTING CONTROL SYSTEM

The linearized and beam-pointing control system in Figure 100 of Chapter 4 can be redrawn as shown in Figure 4-1, in which the variable s is the Laplace transform variable of the continuous time t. R(s) is the transformed beacon image position. $C_1(s)$ and $C_2(s)$ are, respectively, the position of the tracking telescope optical axis and the position of the tracking transfer lens when referred to the optical axis. C(s) is the transfer lens to Earth beacon line of sight. $K_1(s)$, $K_2(s)$ and $G_1(s)$, $G_2(s)$ are the respective controllers and transfer functions.

The condition of stability of the individual loops is that the characteristic equations

$$1 + K_1(s)\,G_1(s) = 0$$

$$1 + K_2(s)\,G_2(s) = 0$$

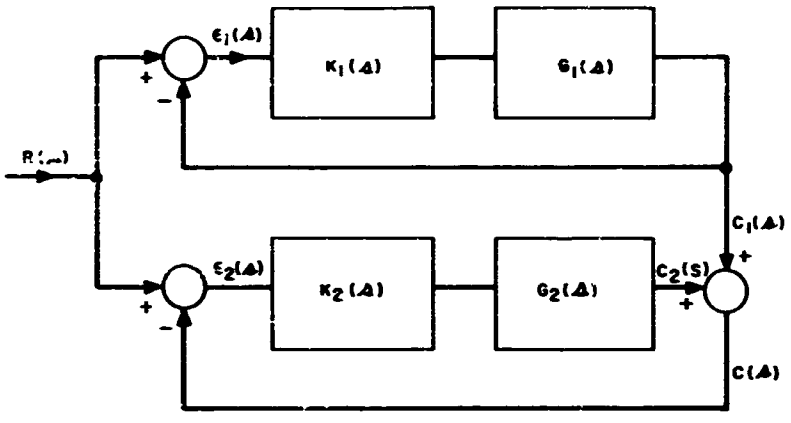
must have no roots in the right half of the s plane.

The characteristic equation of the composite system is

$$[1 + K_1(s)\,G_1(s)]\,[1 + K_2(s)\,G_2(s)] = 0$$

Clearly, the stability of the individual loops implies the stability of the composite system.

When the beacon signal is pulsed, the requirements for stability translate to the condition that the discrete version of the characteristic equations of the individual loops shall have no roots with magnitude equal to or greater than unity.



FOR THE INDIVIDUAL LOOPS

$$\frac{C_1(\Delta)}{R(\Delta)} = \frac{K_1(\Delta)\,G_1(\Delta)}{1 + K_1(\Delta)\,G_1(\Delta)}$$

$$\frac{C_2(\Delta)}{R(\Delta)} = \frac{K_2(\Delta)\,G_2(\Delta)}{1 + K_2(\Delta)G_2(\Delta)} \qquad C_1(S) = 0$$

FOR THE COMPOSITE SYSTEM

$$\frac{C(\Delta)}{R(\Delta)} = \frac{K_1(\Delta)\,G_1(\Delta) + (1 + K_1(\Delta)\,G_1(\Delta))\,K_2(\Delta)\,G_2(\Delta)}{(1 + K_1(\Delta)\,G_1(\Delta))\,(1 + K_2(\Delta)\,G_2(S)}$$

Figure 4-1. Stabiiity of beam pointing control system

APPENDIX 5. THE GRAND LOOP

It was shown in Section 3.2 of Chapter 4 that the control and stability problem associated with the grand loop during acquisition can be treated as a problem involving a dynamical system with constant time delay between the input and the output. Since error detection for the grand loop is discrete in time, the controller design can best be treated in the discrete time domain using techniques discussed in R. W. Koepcke, "On the Control of Linear Systems with Pure Time Delay," JACC (1964).

1. SYSTEM MODEL

The coordinate system chosen for the beam pointing control system on the space vehicle is defined in Figure 5-1. In this figure the x_1x_2 plane is perpendicular to the apparent line of sight (ALOS) from the vehicle telescope to the Earth station. The x_1 axis is the projection of the vehicle – Sun line into the x_1x_2 plane, and the x_2 axis is the line perpendicular to x_1 and the ALOS. The nominal pointing direction x_N is obtained initially from the target ephemeris. A sequence of scanning directions x_S is superimposed on x_N. The scanning beam scans with a suitable beam size through a given field of search around the nominal pointing direction x_N. It carries codes identifying its position with respect to x_N. When a reception of the scanning beam is made at the Earth station, a pointing error relative to x_N can be identified by the code. Because of the finite beam size of the scanning beam, the measured pointing error is quantized. This error is used to generate pointing corrections, which can either be calculated at the Earth station and then sent to the space vehicle or calculated aboard the space vehicle after the measured error is sent to the space vehicle from the Earth station.

A block diagram illustrating the operations described above is shown in Figure 5-2. In this figure $x = x_N + x_S$ is the actual pointing direction and y is the same pointing direction but as seen at the Earth station. A pointing error z results from a difference between the nominal beam position y_N (i.e., the delayed value of x_N) and the position of the Earth station receiver y_T. If the correction signal is generated on the space vehicle, a delayed value of z represented by a new variable, e, is received by the space vehicle. Based on e and the point-ahead signal a new control signal u is generated.

In view of the long time delay existing in the grand loop, the dynamics of the beam steering mechanism will generally be negligible. Also, because of the long-term delay, the ability of the controller to regulate against short-term disturbances and errors is severly limited. (These will be taken out by the beacon tracking and attitude control loops.) Only biasing and calibration errors and long-term drifts can be accounted for by the controller. Thus it is appropriate to describe the grand loop by the following system of equations:

$$\begin{aligned} x(k) &= Gu(k) + b(k) + n_1(k) \\ b(k) &= b(k-1) + n_2(k-1) \\ y(k) &= x(k-p) \\ z(k) &= y(k) - y_T(k) + n_3(k) \\ e(k) &= z(k-p) + n_4(k) \end{aligned} \tag{1}$$

where k is the running variable indicating the k^{th} sampling instant $t_k = k\Delta t$, Δt being the sample interval; n_1 represents the input noise to the beam steerer, n_2 the internal noise accounting for the long-term drift, n_3 the measurement and quantization noise, and n_4 the additional measurement noise at the space vehicle; b represents the unknown biases within the grand loop; G represents the gain of the beam steerer; and $p = T_d/\Delta t$ accounts for the one-way delay time. Each noise component is assumed to have a zero mean, to be uncorrelated from sample to sample, and to be uncorrelated with other noise components. Recursive substitution of the equations in Equation (1) gives

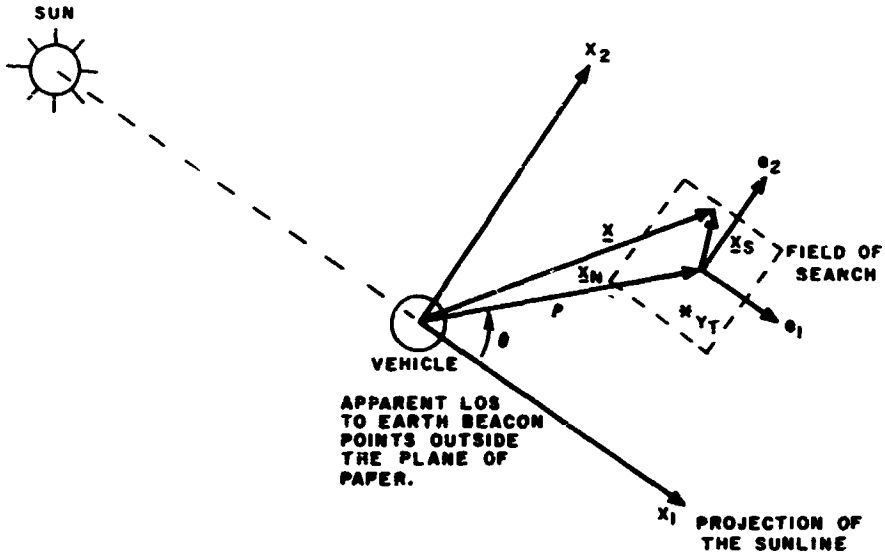
$$\begin{aligned} e(k) = {} & Gu(k-2p) + b(k-2p) + \Big(n_1(k-2p) + n_3(k-p) \\ & + n_4(k)\Big) - y_T(k-p) \end{aligned} \tag{2}$$

Taking the difference of the error e(k) between successive sampling instants gives the following expression for error propagation:

$$e(k) = e(k-1) + G\Delta u(k-2p) - \Delta y_T(k-p) + v(k-1)$$

where

$$\begin{aligned} \Delta u(k-2p) &= u(k-2p) - u(k-2p-1) \\ \Delta y_T(k-p) &= y_T(k-p) - y_T(k-p-1) \\ v(k-1) &= \Big[n_2(k-2p-1) + n_1(k-2p) \\ &\quad - n_1(k-2p-1) + n_3(k-p) \\ &\quad - n_3(k-p-1) + n_4(k) - n_4(k-1)\Big] \end{aligned} \tag{3}$$



(ρ, θ) OBTAINED FROM TARGET EPHEMERIS

$\underline{Y}_T$ = ACTUAL TARGET POSITION

$\underline{X}_N$ = NOMINAL POINT AHEAD

$\underline{X}_S$ = SCANNING BEAM DIRECTION

$\underline{X}$ = ACTUAL POINTING DIRECTION

Figure 5-1. Coordinates for the grand loop control

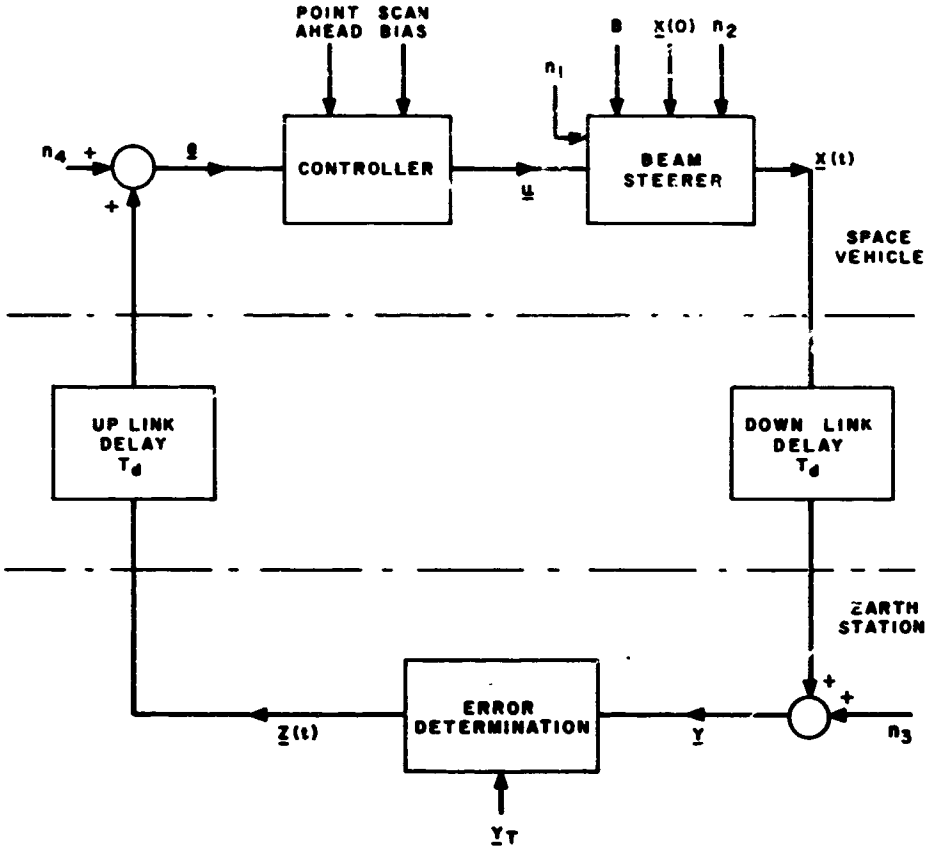


Figure 5-2. System configuration for the grand loop

It is seen that v(k) has zero mean, but it is correlated from sample to sample. However, for small values of p, i.e., large sampling intervals, the correlation can be neglected.

2. CONTROLLER DESIGN

The motion of the Earth station receiver $\Delta y_T(k)$ in Equation (3) is compensated for by the point-ahead signal. Thus, if the presence of the noise term v(k-1) in Equation (3) is neglected, the equation for error propagation becomes

$$e(k) = e(k-1) + G\Delta u(k-2p) \tag{4}$$

The system is designed so that the closed loop error propagation can be expressed as

$$e(k) = \beta e(k-1) \tag{5}$$

with $|\beta| < 1$ for stable operation. Thus the controller output must be

$$\Delta u(k) = C \cdot e(k+2p-1) = G^{-1}(\beta - 1) \cdot e(k+2p-1)$$

$$= C\left(e(k) + G\sum_{i=1}^{2p-1} \Delta u(k-i)\right) \tag{6}$$

and

$$u(k) = u(k-1) + \Delta u(k) = \sum_{j=-\infty}^{k} \Delta u(j)$$

The control action at the k^{th} sampling time is proportional to the sum of the error at the k^{th} sampling time and the effect of the previous control actions during one round trip time delay. By adding the effect of previous control actions, overcorrection of the error is avoided.

3. INCORPORATION OF THE POINT-AHEAD INTO THE CONTROLLER

The angular motion of the Earth station receiver relative to the space vehicle is a time-varying function. However, as shown in Figure 99 of Chapter 4, at Mars distances the angular rate appears quite small. It will be adequate to consider only linear motion, that is, motion in which

$$y_T(k) = a_0 + (a_1\Delta t)k \tag{7}$$

where a_0 is the relative position
a_1 is the relative velocity
$\Delta t = T_d/P$ is the sample time

The point-ahead signal for this case is

$$u_1(k) = G^{-1}\left(a_0' + a_1'(k+p)\Delta t\right)$$

$$= G^{-1}(a_0' + a_1' p\Delta t) + (G^{-1}a_1'\Delta t)k \tag{8}$$

where u_1 is the point-ahead part of the controller output

a_0' and a_1' are the estimated target position and velocity

In equation (8), the first term on the righthand side can appear as an initial condition for the control u(k) in Equation (6), while the second term modifies Equation (6) so that

$$\Delta u(k) = C\left[e(k) + G\sum_{i=1}^{2p-1} \Delta u(k-i)\right]$$

and $u(k) = u(k-1) + \Delta u(k) + G^{-1}a_1'\Delta t$ (9)

with $u(0) = G^{-1}(a_0' + a_1' p\Delta t)$

Here it has been assumed that point-ahead is initiated at the first sampling instant. Since both a_1' and a_1 are small, the angular rate a_1' can be neglected during the acquisition phase. The error introduced can be readily estimated from Equations (3) and (5) by noting $\Delta y_T(k) = a_1\Delta t$. From these equations, the steady-state error is $e_{ss} = (1-\beta)^{-1} a_1 \Delta t$, which is the analog of the velocity error in conventional servo theory.

4. THE CONTROLLER AS A FILTER

Since the controller is basically the discrete version of an integrator, certain filtering action can be introduced into the control loop by suitably increasing the sampling rate. This can be shown by considering the one-dimensional case as follows: Suppose it is required that in the noiseless case a pointing error $\hat{e}(0)$, which has been present till time t = 0, be reduced by time T to $\rho e(0)$, $0 < \rho < 1$. If there are N sampling instants in the time period [0, T], then designing the system so that $\hat{e}(k) = \beta\hat{e}(k-1)$, where $\beta^N = \rho$, will meet the same requirement. It is apparent from Equation (6) that this can be achieved by letting

$$\Delta u(k) = G^{-1}(\beta - 1)\left(e(k) + G\sum_{i=1}^{2p-1}\Delta u(k-i)\right) \qquad (10)$$

In the case where noise is present, the measured error, e, can be expressed in terms of the sum of the true error $\hat{e}$ and the noise by the equation

$$e(k) = \hat{e}(k-2p) + v(k),$$

where v(k) represents the noise component. The round trip delay between the true error e and the measured error $\hat{e}$ is accounted for by setting

$$p = NT_d/T$$

Equation (10) can be rewritten as

$$\Delta u(k) = G^{-1}(\beta-1)\left(\hat{e}(k-2p) + v(k) + G\sum_{i=1}^{2p-1}\Delta u(k-i)\right) \qquad (11)$$

Since the control action begins at k = 1, and $\hat{e}(k) = \hat{e}(0)$ for $k \leq 0$, the general form of $\Delta u(k)$ is

$$\Delta u(k) = G^{-1}(\beta-1)\left[\beta^{k-1}\hat{e}(0) + \left((\beta-1)\sum_{i=1}^{k-1}\beta^{k-i-1}v(i)\right) + v(k)\right] \qquad (12)$$

Consequently,

$$u(N) - u(0) = G^{-1}(\beta-1)\left[\sum_{k=1}^{N}\left\{\beta^k\hat{e}(0) + (\beta-1)\sum_{i=1}^{k-1}\beta^{k-i-1}v(i) + v(k)\right\}\right]$$

$$= G^{-1}(\beta-1)\left[\left(\frac{1-\beta^N}{1-\beta}\right)\hat{e}(0) + \sum_{k=1}^{N}\frac{(\beta-1)(1-\beta^k)-(\beta-1)}{(1-\beta)}v(k)\right]$$

$$= G^{-1}\left[(\beta^N - 1)\hat{e}(0) + (\beta-1)\sum_{k=0}^{N-1}\beta^k v(N-k)\right] \qquad (13)$$

or

$$\hat{e}(N) = \beta^N\hat{e}(0) + (\beta-1)\sum_{k=0}^{N-1}\beta^k v(N-k). \qquad (14)$$

Now assume the noise components v(k) are uncorrelated and that each has a standard deviation σ_v. Then, letting σ_N denote the standard deviation of $\hat{e}(N)$, one obtains

$$\sigma_N = \sigma_v(1-\beta)\sqrt{\frac{1-\beta^{2N}}{1-\beta^2}}$$

$$= \sigma_v\sqrt{\frac{(1-\beta^{2N})(1-\beta)}{(1+\beta)}} \qquad (15)$$

or

$$\sigma_N = \sigma_v\sqrt{\frac{(1-\rho^2)(1-\rho^{1/N})}{1+\rho^{1/N}}} \qquad (16)$$

For $0 < \rho < 1$, the expression $(1-\rho^{1/N})/(1+\rho^{1/N})$ is a decreasing function of N. Consequently, increasing the sampling rate (accompanied by a corresponding increase in β, that is a corresponding decrease in the controller gain) decreases the variances of the output.

If the noise components are correlated from sample to sample, one cannot obtain as simple an expression as Equation (16). However, when the correlation time is much shorter than the sample time, the above argument remains qualitatively correct.

It should be noted that the filtering action is obtained at the expense of additional memory storage required for storing the previous control action during the $T_d = 2T_d$ delay time and at the expense of more frequent controller motion. Of course, a separate filter can be used if the higher frequency of the controller action is causing undesirable actuator wear.

5. EXAMPLE

To illustrate the effect of different controller gain settings and the effect of the different sample rates, consider a simple one-dimensional example (the motions x_1x_2 are decoupled) corresponding to a situation in which the point-ahead information is available, but a biasing error exists in the grand loop. This biasing error can be caused either by the long-term component drifts and the

calibration errors or by the inaccuracy in the point ahead information. For convenience, per unit quantities are used in Equations (1), (7), and (9), so that the system parameters are

$$G = 1.0, \quad b = 1.0, \quad a_0 = a_0', \quad a_1 = a_1'$$

(for perfect point ahead) and p = 1, and the initial error is -1.0.

The error responses for control settings C = -1.0, -0.9, -0.75, -0.5, -0.25 for the noiseless case ($n_1 = n_2 = n_3 = n_4 = 0$) are shown in Figure 5-3. It is interesting to note that the closed-loop response with a conventional discrete integrating error compensation [i.e., $\Delta u(k) = Ce(k)$] is unstable for all the above values of C with the exception of C = -0.25. Thus, by considering the delay time effect in the controller design, a much faster response time can be achieved (compare the responses of Figure 5-3a and 5-3e). The point-ahead is assumed to be initiated at t = 0, but no error is indicated on the space vehicle until an error detection is made at the Earth station and returned to the space vehicle. Thus, in Figure 5-3 an error indication first appears at $t = 2T_d$ (T_d being the one-way delay time). (For p = 1, T_d is also the sampling interval.) This error indication does not change until the time $t = 4T_d$ which corresponds to one round-trip delay from the time of initiation of control action. This initial error response will be independent of the controller settings. Subsequent error response, however, is a function of the controller setting C as shown in Figure 5-3. Thus the total error response time can be separated into two parts; a gain independent part $\overline{T}_d = 2T_d$, corresponding to one round-trip delay time and a gain dependent part T_c, which is the time interval between the time when error begins to change and the time when the error becomes zero ($|e| \leq \epsilon$, ϵ predetermined). In Figure 5-3, case (a), C = -1.0, $\overline{T}_d = 2T_d$, $T_c = 0$, and hence the total response time is $T = 2T_d$. In case (e), C = -0.25, $\overline{T}_d = 2T_d$, $T_c = 13T_d$, and hence the total response time is $T = 15T_d$.

The effect of increasing the sampling rate is illustrated in Figure 5-4, where a 10 percent (compared with $|e_{max}| = 1.0$) measurement noise which is uniformly distributed between -0.1 and 0.1 is assumed to occur at the Earth station terminal. A comparison among cases (a), (b), and (c) in Figure 5-4 shows the effective increase in loop gain by noting that the error response for case (b), where p = 6 and C = -0.25, is essentially the same as case (c), where p = 1 and C = -0.75. A comparison between cases (b) and (d) in Figure 5-4 shows the less pronounced noise response (i.e., the filtering action) in case (b) because of a decrease in controller setting from C = -0.75 to C = -0.25, while the equivalent loop gain remains unchanged by increasing the sampling rate from p = 1 to p = 6.

Figure 5-5 and 5-6 illustrate system error responses in the presence of both bias error and noise. The bias error is the same as above. The noise components are either uniformly distributed between -0.1 and 0.1 or are Gaussian-distributed with zero mean and $\sigma = 0.05$. An error response behavior which is similar to that in Figure 5-4 can be observed in these figures.

The controller setting C and the sampling rate p must be chosen compositely to (1) achieve desired response to system errors and (2) limit the control action because of the presence of noise (thus providing adequate filtering action).

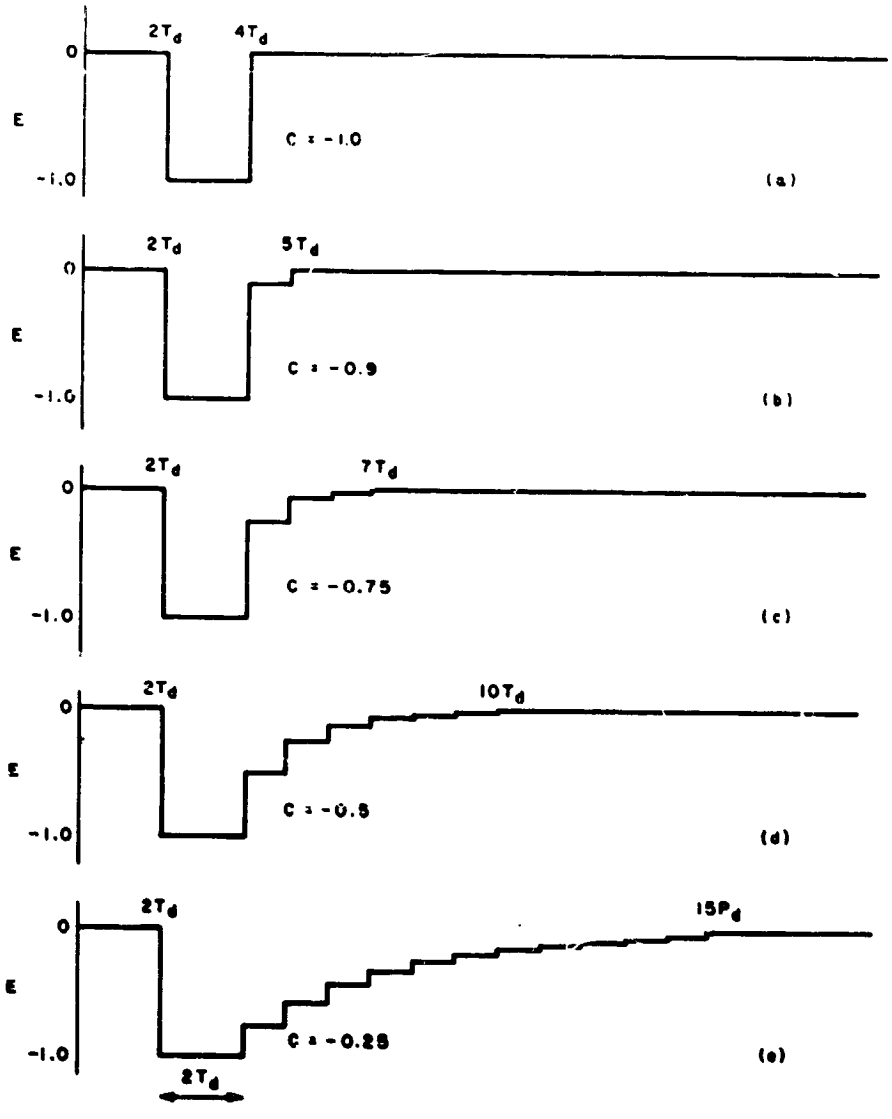


Figure 5-3. Error responses for various gain settings – noiseless case

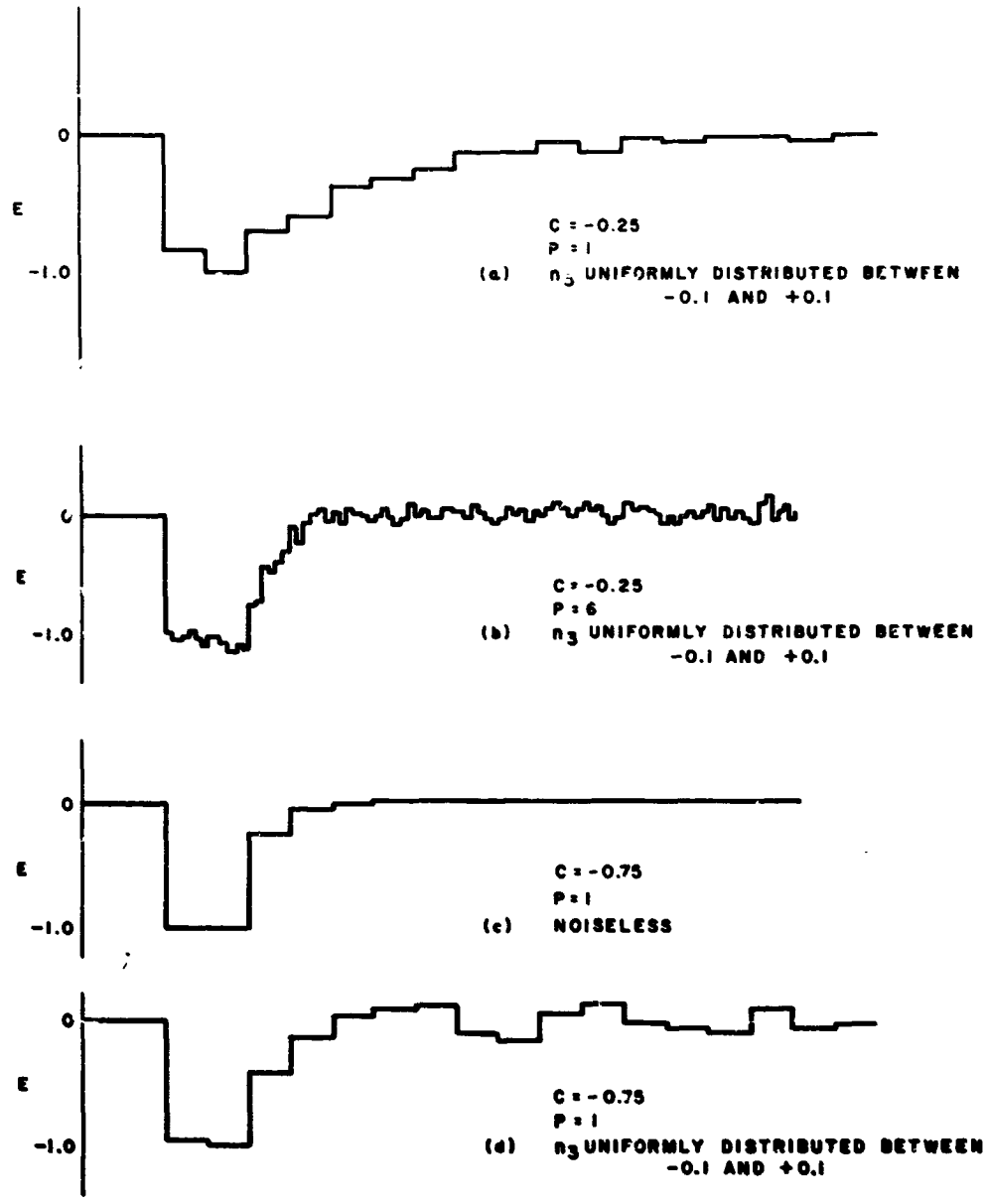


Figure 5-4. Error responses for different sampling rates and gain settings

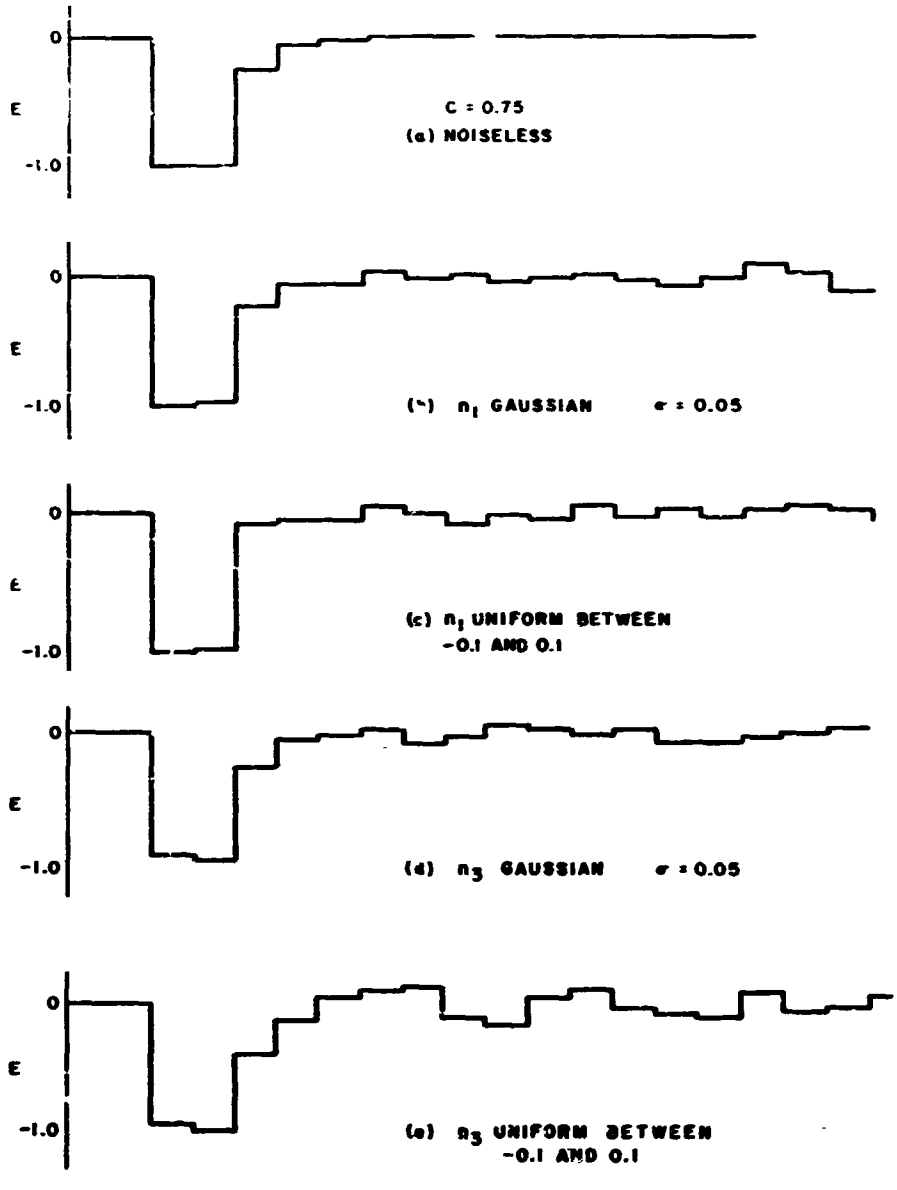


Figure 5-5. Error response for C = -0.75 and p = i.

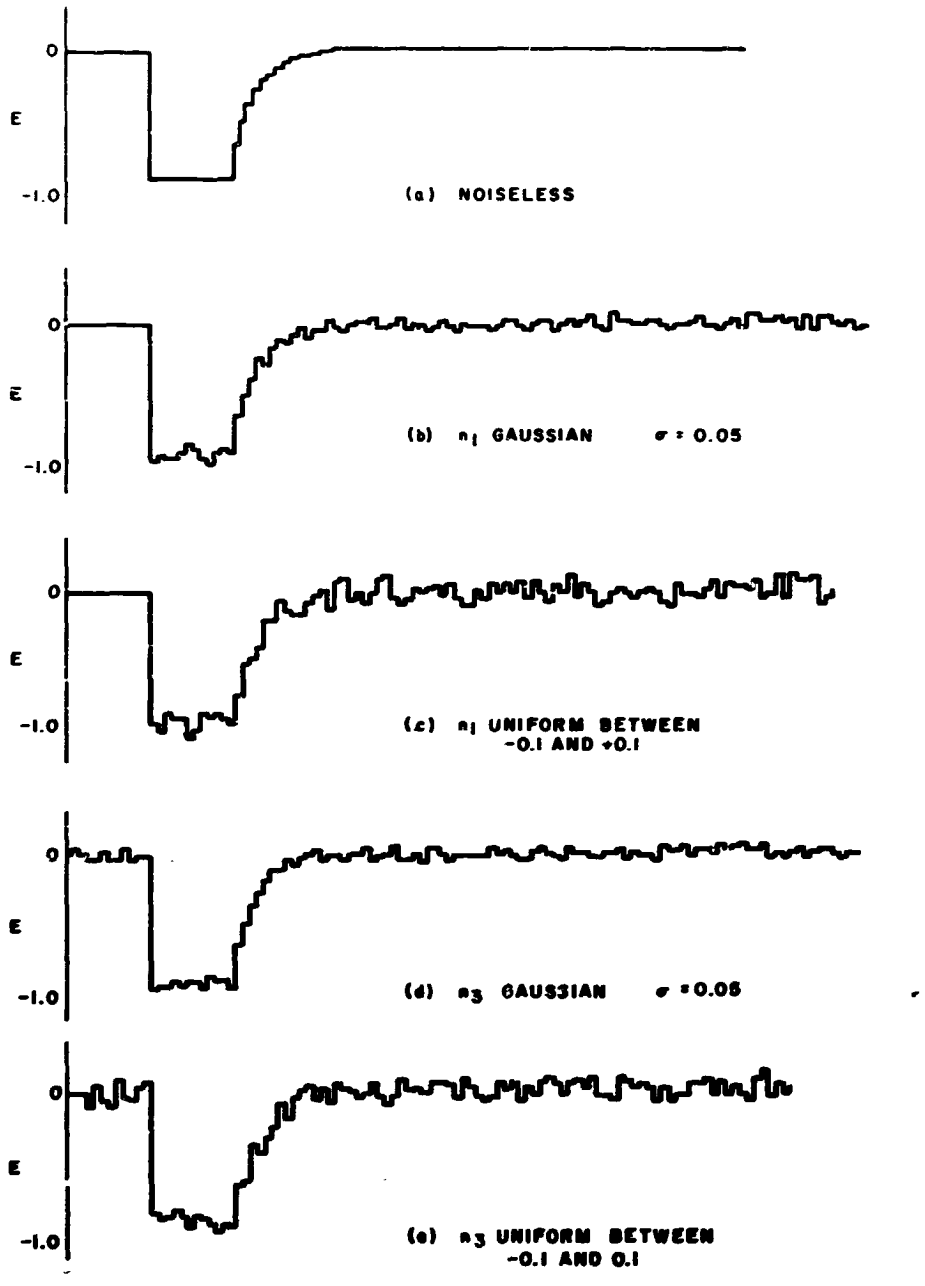


Figure 5-6. Error responses for C = -0.25 and p = 6.

APPENDIX 6.

HETERODYNE COHERENCE AREA AND ANGULAR ALIGNMENT

The general expression for $\bar{i}_{so}^2$ was given in Section 6.3 of Chapter 4, Equation (32). This relation was obtained from the knowledge that the photocurrent is related to the aperture electric field by

$$i_{phot} = \int_A \frac{\eta\tau_o e}{h\nu} \frac{E^2}{Z} d^2\vec{r} \qquad (1)$$

where E is the total instantaneous electric field at $\vec{r}$. The total aperture field of course is a superposition of signal, background, and local fields. Assume that each of these fields can be written as

$$E_p(\vec{r}, t) = x_p(\vec{r}, t) \cos \omega_p t + y_p(\vec{r}, t) \sin \omega_p t \qquad (2)$$

(p = s,b,o for signal, background, and local, respectively) and that x and y are statistically identical, but independent, random variables with zero means. Writing the total electric field as a superposition of the three component fields, the sum may be squared and inserted in Equation (1) to yield i_{phot}. Among the components of i_{phot} is i_{so}, which is expressed by

$$i_{so} = \left(\frac{\eta\tau_o e}{h\nu Z}\right) \int_A \Big| (x_s x_o + y_s y_o) \cos(\omega_s - \omega_o) t + (y_s x_o - x_s y_o) \sin(\omega_s - \omega_o) t \Big| d^2\vec{r} \qquad (3)$$

Squaring and retaining only the expected value of the time-averaged portion of i_{so}^2, leads to Equation (32) in Chapter 4:

$$\bar{i}_{so}^2 = 2\left(\frac{\eta\tau_o e}{h\nu Z}\right)^2 \int_A \int_A \overline{x_s x_s'}\ \overline{x_o x_o'}\ d^2\vec{r}\, d^2\vec{r}'$$

[Ch 4. Eq. (32)]

Analogous forms describe the remaining mixing currents (note that $\bar{i}_{so}^2$, etc., denote time-averaged, mean-square currents). When the fields display perfect spatial coherence and have uniform intensity distributions, Equation (32) is easily evaluated and yields

$$\bar{i}_{so}^2 = 2\left(\frac{\eta\tau_o e}{h\nu}\right)^2 P_s P_o$$

as noted earlier.

Consider now the case in which the beam intensities have Gaussian distributions:

$$P_s(\vec{r}) = \frac{\overline{x_s^2(\vec{r})}}{Z} = P_{sm} e^{-r^2/R_s^2}, P_o(\vec{r}) = \frac{\overline{x_o^2(\vec{r})}}{Z} = P_{om} e^{-r^2/R_o^2}$$

Here P_{sm} and P_{om} are the expected peak-power densities at the beam centers. The correlations within the beams are also assumed to be Gaussian functions

$$\frac{\overline{x_s x_s'}}{Z} = P_s(\vec{r}) e^{-(\Delta r)^2/\rho_s^2}, \frac{\overline{x_o x_o'}}{Z} = P_o(\vec{r}) e^{-(\Delta r)^2/\rho_o^2}$$

where $\Delta r = |\vec{r} - \vec{r}'|$, and the coherence areas $a_s = \pi\rho_s^2$ and $a_o = \pi\rho_o^2$ are $\ll A_R$ (or either of the beam areas $A_o = \pi R_o^2$, $A_s = \pi R_s^2$). Hence variations in the power densities over distances of ρ_s and ρ_o are neglected. [Otherwise, in defining the coherence functions, $P_s(\vec{r})$ would have to be replaced by $P_s(\vec{r})P_s(\vec{r}')$, or some other suitable function of $P_s(\vec{r})$ and $P_s(\vec{r}')$.]

With the foregoing assumptions, $\overline{i_{so}^2}$ may again be evaluated in a straightforward manner to yield

$$\overline{i_{so}^2} = 2\left(\frac{\eta\tau_o e}{h\nu}\right)^2 P_{sm}P_{om}\left(1 - e^{-A_R/A_r}\right)\frac{a_s a_o}{a_s + a_o}A_r \quad (4)$$

where $A_r = \dfrac{A_s A_o}{A_s + A_o}$. From Equation (33) in Chapter 4 it can be seen that in this situation

$$F = \left(1 - e^{-A_R/A_r}\right)\frac{\max(a_s,a_o)}{a_s + a_o}\frac{A_r}{A_R}$$

The calculation here is simplified by the fact that the minimum coherence area is $<<A_R$. When the coherence areas are smaller than, but not negligible compared to, A_R, evaluating Equation (32) is complicated by the necessity for including edge effects at the photosurface boundaries.

A more general form for i_{so}^2 which includes the effects of misaligned signal and local fields is

$$i_{so}^2 = 2(\cos\alpha)\left(\frac{\eta\tau_o e}{h\nu Z}\right)^2 \iint \overline{x_s(\vec{r},t)\,x_s(\vec{r}',t)}\;\overline{x_o(\vec{r},t)\,x_o(\vec{r}',t)}\;\cos\vec{k}_s\,(\vec{r}' - \vec{r})\,d^2\vec{r}\,d^2\vec{r}' \quad (5)$$

This expression is appropriate when the local field impinges normally on the photosurface and the signal field approaches at an angle α from the normal. $\vec{k}_s$ is the nominal signal-wave vector and the vectors $\vec{r}$ and $\vec{r}'$ lie in the plane of the photosurface. When ideal monochromatic plane wave fields are assumed, the angular distance to the first null in Equation (5) may be used to define Ω_R, in which case it is well known that

$$\Omega_R = \frac{\lambda^2}{A_R} \qquad \text{[Ch 4. Eq. (34)]}$$

In the ideal case, nulls in $\overline{i_{so}^2}$ result from destructive interference among the incremental current contributions, which in turn arise from the relative shift in carrier phase across the full aperture. When the fields are not spatially coherent, it can be assumed that the relative shift in carrier phase over the smallest coherence area determines the first null in $\overline{i_{so}^2}$, with the result that Ω_R is given by

$$\Omega_R = \frac{\lambda^2}{\min[a_s,a_o]} \quad (6)$$

Hence the ratio of the field of view of the partially coherent heterodyne to that of the completely coherent heterodyne is

$$\frac{\Omega_{pc}}{\Omega_{coh}} = \frac{A_R}{\min[a_s,a_o]} \quad (7)$$

The dependence of heterodyne performance on coherence area may also be understood from a slightly different (quantum-mechanical) viewpoint. As just noted, two plane waves with propagation vectors making an angle greater than about λ/d do not interfere, on the average, over an aperture with linear dimension d. That is, waves in distinguishably different transverse modes (or photons in different cells of phase space) do not interfere.[2] Thus, the area over which plane waves may interfere or the coherence area, a_c, corresponds to an angular spread Ω of propagation vector for which constructive interference takes place $a_c = \lambda^2/\Omega$. The turbulent behavior of the atmosphere changes (spreads out) the apparent direction of propagation of the incoming wave. To calculate the signal and shot noise in a heterodyne receiver, one must take into account the number of spatial modes over which the signal and the local oscillator waves are distributed. Suppose, then, that in a given time interval one has $\underline{s}$ signal photons and $\underline{o}$ local oscillator photons evenly distributed over N_s, N_o modes respectively, and that these modes are spread out about a mean direction of propagation which is the same for the two waves. The mixing term, representing the communication signal, is proportional to

$$\sum_k \delta_{sk}\delta_{ok}$$

where $\delta_{sk} = s/N_s$, $\delta_{ok} = o/N_o$, and the summation is carried out over all the modes which contain both signal and LO photons. Thus, the mixing term is

$$\frac{\min(N_o,N_s)}{N_o N_s}\,s\cdot o,$$

and the number of photoelectrons within the IF band is

$$\left(\eta\tau_o\right)^2\frac{\min(N_o,N_s)}{N_o N_s}\,s\cdot o$$

The shot noise is just

$$\eta\pi_o\left[\sum_{N_o}\delta_{ok}+\sum_{N_s}\delta_{sk}\right] = \eta\pi_o(o+s) \approx \eta\pi_o \cdot o$$

assuming the local oscillator is much stronger than the signal. The resulting signal-to-noise ratio is

$$snr = \left(\frac{\eta\pi_o}{h\nu}\right)\frac{\min(N_o,N_s)}{N_oN_s}\pi_sA_Pt$$

where t is the observation interval. Now

$$N_o = \frac{A_R}{\lambda^2/\Omega_o} = \frac{A_R}{a_o}$$

where Ω_o, a_o represent the apparent solid angle and corresponding coherence area for the local oscillator. This is only the case, however, if $A_R > a_o$, since if $A_R < a_o$ the local oscillator cannot be resolved by the receiver. Hence

$$N_o = \begin{cases} A_R/a_o & \text{if } A_R > a_o \\ 1 & \text{if } A_R < a_o \end{cases}$$

and likewise

$$N_s = \begin{cases} A_R/a_s & \text{if } A_R > a_s \\ 1 & \text{if } A_R < a_s \end{cases}$$

This leads to the following results:

$$snr = \frac{\eta\pi_o}{h\nu}\pi_st \cdot \left\{\begin{array}{l} A_R \text{ if } A_R < a_s < a_o \\ a_s \text{ if } a_s < a_o, A_R \\ A_R \text{ if } A_R < a_o < a_s \\ a_o \text{ if } a_o < a_s, A_R \end{array}\right\} =$$

$$\min[A_R, a_s, a_o] \qquad (8)$$

This is the result embodied in Equation (33) of Chapter 4.

APPENDIX 7. HOMODYNE INFORMATION RATE

Considering the simple polarization modulation scheme (Figure 108, Chapter 4), the signal and LO fields at the receiving aperture may be written in the form

$$V_s(\vec{r}, t) = X_s(\vec{r})\sin[\omega_s t + \phi_s(\vec{r})] \quad (1)$$

$$V_o(\vec{r}, t) = X_o(\vec{r})\sin[\omega_o t + \phi_o(\vec{r})] \quad (2)$$

The time dependence of X and ϕ is suppressed, since these quantities do not vary over a bit period (at least in high data rate systems). Similarly the background field is expressed as

$$V_b(\vec{r}, t) = X_b(\vec{r}, t)\sin[\omega_b t + \phi_b(\vec{r}, t)] \quad (3)$$

In Equation (3), the variation of amplitude and phase over a bit period cannot be ignored.

The output of the PMT in the channel containing the signal is

$$i_a = M(i_s+i_o+i_b+i_d+i_{so}+i_{sb}+i_{bo}+i_{ns}+i_{nb} + i_{no}+i_{nd}+i_{th/M}) \quad (4)$$

Expressions for typical direct detected and mixed currents; e.g., i_o and i_{so}, appear as

$$i_o = \left(\frac{\eta\tau_o e}{h\nu}\right)\int_A \frac{X_o^2(\vec{r})}{2Z}\,d^2\vec{r} \quad (5)$$

$$i_{so} = \left(\frac{\eta\tau_o e}{h\nu}\right)\int_A \frac{X_s(\vec{r})X_o(\vec{r})}{Z}\cos[\phi_s(\vec{r}) - \phi_o(\vec{r})]\,d^2\vec{r} \quad (6)$$

where it is assumed that the LO laser is phase-locked to the incoming signal ($\omega_s - \omega_o = 0$).

Assuming a strong LO, Equation (4) becomes

$$i_a \simeq i_o + i_{so} + i_{bo} + i_{no} \quad (7)$$

The integrator output I_a^2 is obtained by squaring Equation (7) and integrating, with the result

$$I_a^2 = (i_o + i_{so})^2 + 2(i_o + i_{so})\frac{1}{T}\int_0^T (i_{bo} + i_{no})dt + \frac{1}{T}\int_0^T (i_{bo} + i_{no})^2 dt \quad (8)$$

From Equation (8) one finds

$$\overline{I_a^2} = (i_o + i_{so})^2 + \frac{1}{T}\int_0^T \left(\overline{i_{bo}^2} + \overline{i_{no}^2}\right)dt \quad (9)$$

An analogous procedure yields $\overline{I_\beta^2}$, and taking the difference,

$$\overline{\Delta I^2} = \overline{I_a^2} - \overline{I_\beta^2} = 2i_o i_{so} + i_{so}^2 \quad (10)$$

Computing the variance of I_a^2 and I_β^2 is slightly more involved, but straightforward. The result for I_a^2 is

$$\sigma^2_{I_a^2} = \frac{4}{T^2}\left(i_o^2 + 2i_o i_{so} + i_{so}^2\right)\int_0^T\int_0^T \left(\overline{i_{bo}i'_{bo}} + \overline{i_{no}i'_{no}}\right)dtdt' + \frac{4}{T^2}\int_0^T\int_0^T \overline{i_{bo}i'_{bo}}\;\overline{i_{no}i'_{no}}\,dtdt' + \frac{1}{T^2}\int_0^T\int_0^T \left(\overline{i_{bo}^2 i_{bo}'^2} - \overline{i_{bo}^2}\;\overline{i_{bo}'^2} + \overline{i_{no}^2 i_{no}'^2} - \overline{i_{no}^2}\;\overline{i_{no}'^2}\right)dtdt' \quad (11)$$

The expression for $\sigma^2_{\overline{I^2}\rho}$ may be obtained by setting $i_{so} = 0$ (since it is deterministic) in Equation (11). The sum of variances then follows directly from Equation (11). Employing the Gaussian assumption for the integrator outputs leads as before to

$$\frac{\overline{\Delta I^2}^2}{\sigma^2_{\Delta I^2}} = K(P_e) \quad (12)$$

which, after substituting the foregoing expressions and minor rewriting, yields an expression for the information rate H:

$$H = \frac{\frac{1}{K}\left(i_o^2 + i_o i_{so} + \frac{i_{so}^2}{4}\right)i_{so}^2}{\left[(2i_o^2 + 2i_o i_{so} + i_{so}^2)\frac{1}{T}\int_0^T\int_0^T \left(\overline{i_{bo}i'_{bo}} + \overline{i_{no}i'_{no}}\right)dtdt' + \frac{2}{T}\int_0^T\int_0^T \overline{i_{bo}i'_{bo}}\;\overline{i_{no}i'_{no}}\,dtdt' + \frac{1}{2T}\int_0^T\int_0^T \left(\overline{i_{bo}^2 i_{bo}'^2} - \overline{i_{bo}^2}\;\overline{i_{bo}'^2} + \overline{i_{no}^2 i_{no}'^2} - \overline{i_{no}^2}\;\overline{i_{no}'^2}\right)dtdt'\right]} \quad (13)$$

For a sufficiently powerful local oscillator, the contribution from the heterodyned background field is dominated by LO shot noise. Under these conditions, Equation (13) may be written

$$H \sim \left(\frac{1}{2K}\right) \frac{i_{so}^2}{\frac{1}{T}\int_0^T \int_0^T \overline{i_{no} i'_{no}}\, dt dt'} \sim \left(\frac{B}{K}\right) \frac{i_{so}^2}{i_{no}^2} \qquad (14)$$

Substituting for the indicated currents, one finds

$$H \sim \left(\frac{B}{K}\right) \frac{4\left(\frac{\eta\tau_o e}{h\nu}\right)^2 P_s P_o}{2e\left(\frac{\eta\tau_o e}{h\nu}\right) P_o B} = \frac{2}{K}\left(\frac{\eta\tau_o P_s}{h\nu}\right) \qquad (15)$$

APPENDIX 8.

MIXING-TERM FLUCTUATIONS FOR DIRECT DETECTION

The classical limit for direct detection, in which mixing terms dominate shot-noise terms, is illustrated in this appendix. The signal field is assumed to have constant magnitude, a, and to lie in a single spatial mode. The noise field is characterized by orthogonal Gaussian random variables, $X_{\kappa\ell}$, $Y_{\kappa\ell}$ (mean zero, variance σ^2), where κ denotes the temporal mode, $\kappa = 1,2, \ldots, WT(\equiv K)$ and ℓ the spatial mode, $\ell = 1,2, \ldots, \Omega_R A_R/\lambda^2)\ (\equiv L)$.

The time-averaged output current is (omitting inessential constants)

$$I = \frac{1}{T}\int_0^T i\,dt$$

$$\approx \frac{1}{K}\sum_{\kappa=1}^{K} i_\kappa$$

where $i_\kappa = \frac{1}{2}\left|\left(a + X_{\kappa 1}\right)^2 + Y_{\kappa 1}^2 + \sum_{\ell=2}^{L}\left(X_{\kappa\ell}^2 + Y_{\kappa\ell}^2\right)\right|$

As the same is true for all κ, κ may be suppressed, so that

$$\bar{I} = \bar{i}_\kappa$$

and

$$\sigma_I^2 = \bar{I}^2 - \bar{I}^{-2}$$
$$= \frac{1}{K}\sigma_i^2$$
$$= \frac{1}{K}\left[\bar{i}^2 - \bar{i}^{-2}\right]$$

where $i = \frac{1}{2}\left|a^2 + 2X_1 a + \sum_{\ell=1}^{L}\left(X_\ell^2 + Y_\ell^2\right)\right|$

Calculating:

$$\bar{i} = \frac{1}{2}a^2 + L\sigma^2$$

from which one identifies the average instantaneous signal current

$$I_s = \frac{1}{2}a^2$$

and the average instantaneous current due to noise

$$I_n = L\sigma^2$$

Carrying out the indicated expansions, making use of the orthogonality between noise components from different modes, and noting that $\overline{X^4} = \overline{Y^4} = 3\sigma^4$, one obtains

$$\sigma_i^2 = \sigma^2 a^2 + L\sigma^2$$

Thus $\sigma_I^2 = \frac{\sigma^2 a^2}{K} + \frac{L\sigma^4}{K}$

$$= \frac{2(L\sigma^2)(a^2/2)}{KL} + \frac{(L\sigma^2)^2}{KL}$$

$$= \frac{2I_n I_s}{\frac{\Omega_R A_R}{\lambda^2}WT} + \frac{I_n^2}{\frac{\Omega_R A_R}{\lambda^2}WT}$$

which are just the expressions used for σ_3^2, σ_4^2 in the text.

APPENDIX 9. PULSE-POSITION MODULATION

The binary polarization modulation scheme discussed in this appendix requires on the order of $10/\eta\tau_o$ photons per bit of information. Under certain circumstances, it is theoretically possible to design systems (see References 57 and 61 of Chapter 4) which are more efficient in terms of energy use. The capacity of a noise-free optical signal has been well understood since J. P. Gordon's discussion (Reference 7, Chapter 4) of this problem. It was implied that (for unit efficiency, $\eta\tau_o = 1$) on the order of one photon per bit is practical and less than one photon per bit is not proscribed theoretically as long as one is willing to sacrifice large amounts of bandwidth; i.e., when the information rate is small compared with the bandwidth. If ϵ is the number of photons per bit, C the capacity, and B the bandwidth, Gordon's work shows that

$$\frac{C}{B} \approx \frac{1}{\ln 2} \frac{e^{-1/\epsilon}}{\epsilon} \quad (\epsilon < 1)$$

It can be seen that C/B decreases rapidly with decreasing ϵ. If this is not a severe practical restriction, as in the present considerations, one can hope to achieve small values of ϵ. This is in principle what can be done with a PPM system, in which the signal bandwidth increases to produce narrow pulses. A crude analysis of a simple example of such a system, presenting the essentials, follows.

In an interval of time T, a single pulse of duration τ is transmitted in one of $m = T/\tau$ positions. At the receiver, which is assumed to be synchronized, one decides in which of m positions the transmitted pulse occurs, so that the information rate is log m/T bits per second. Assume that the average energy of the transmission can be converted to pulses without loss, so that effectively the signal power in the time interval τ (during transmission) is given by $P_p = P_s T/\tau = P_s m$. If m is large enough, this will be much larger than the background radiation power; hence, it may be possible to reduce P_s at the expense (increased bandwidth) of increasing m. At the same time each pulse contributes log m bits rather than one.

In the PPM system, there are two types of error. In the interval during which the signal pulse is transmitted, it is possible that one fails to detect the signal. In addition, there is a false alarm error when the background noise is large enough to make it appear that the signal is present in some interval when, in fact, it is not. At the receiver, a threshold level is set to decide between signal present or not present. A failed detection error is made when the net signal plus noise fails to exceed the threshold. Assume that the signal is polarized, that thermal noise and dark current can be ignored (i.e., a PMT may be used), and that the output of the time integrator (or base-band filter) is Gaussian. Then considerations in Section 5.7 of Chapter 4 imply that the post-detection integration time T required is given by

$$\frac{1}{\tau} = \frac{1}{k_p} \frac{\dfrac{\eta\tau_o P_s m}{h\nu}}{1 + \dfrac{P_b/2}{P_s m}}$$

where k_p is determined from the desired error probability and the threshold level, and only half the background power produces shot noise since the extraneous polarization may be eliminated at the receiver.

Selection of the threshold value, which in turn will determine k_p for a pre-set error probability, is amenable to analysis* but beyond the scope of this appendix. It will be assumed instead that the threshold is sufficiently high that the total falsealarm probability (for all slots which do not contain the signal pulse) can be neglected. This means that the foregoing relation determines the information rate but that k_p will be somewhat higher than in the binary polarization scheme. The information rate becomes

$$H = \frac{\log m}{T} = \frac{\log m}{ms\tau}$$

or

$$H = \frac{\log m}{k_p} \frac{\dfrac{\eta\tau_o P_s}{h\nu}}{1 + \dfrac{P_b/2}{P_s m}} \qquad (1)$$

*T. Curran and M. Ross, Proc. IEEE, Vol. 53 (1965), p 1770.

One can solve for the signal power required

$$P_s = \frac{k_p h\nu H}{\eta\tau_o \log m}\left\{\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{P_b \cdot \eta\tau_o \log m}{2mk_p h\nu H}}\right\}$$

and, in turn, for the number of photons per bit

$$\epsilon_{PPM} = \frac{P_s}{h\nu H}$$

$$\epsilon_{PPM} = \frac{k_p}{\eta\tau_o \log m}\left\{\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\eta\tau_o P_b}{2k_p h\nu} \cdot \frac{1}{H}\frac{\log m}{m}}\right\} \quad (2)$$

For comparison, from Equation (41) one has

$$P_s = \frac{k_D h\nu H}{\eta\tau_o}\left\{\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\eta\tau_o P_b}{k_D h\nu} \cdot \frac{1}{H}}\right\}$$

for the binary polarization modulation where k_D (D denoting direct detection) is the constant which depends on the desired error probability. The corresponding number of photons per bit is

$$\epsilon_D = \frac{k_D}{\eta\tau_o}\left\{\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\eta\tau_o P_b}{k_D h\nu H} \cdot \frac{1}{H}}\right\} \quad (3)$$

Now forming the ratio for relative power efficiency:

$$\frac{\epsilon_D}{\epsilon_{PPM}} = \log m \left(\frac{k_D}{k_p}\right) \frac{\left\{\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{H^*}{H}}\right\}}{\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{H^*}{H} \cdot \frac{k_D}{2k_p} \cdot \frac{\log m}{m}}} \quad (4)$$

where H^* has been defined by

$$H^* \equiv \frac{1}{k_D}\frac{\eta\tau_o P_b}{h\nu}$$

A detailed comparison of power efficiencies depends, of course, on the number of pulse positions m, but also on the ratio of H^* (which is proportional to the average rate of generation of photoelectrons due to the background) to the information rate H. For a receiver on the ground, H^* may be quite large. For example, if $\lambda = 0.5\mu$, considering a 25 square meter receiver which is atmosphere-limited with Ω_R assumed 10^{-8} sterad, letting $\eta\tau_o = 5 \times 10^{-2}$, taking a 1-angstrom predetection filter and letting $k_D = 10$, one finds $H^* = 2.4 \times 10^7$ (sec^{-1}). Now, suppose $H = 10^6$ bits/second, and assume $k_D \approx k_p$ (as remarked earlier, k_D will usually be somewhat smaller). Thus, Equation (4) becomes approximately

$$\frac{\epsilon_D}{\epsilon_{PPM}} \approx \frac{\log m\sqrt{24}}{\frac{1}{2} + \sqrt{\frac{1}{4} + 12\frac{\log m}{m}}}.$$

The following table shows the power gain (in dB) for the PPM system over the binary system for this illustration. Instead of m, the parameter chosen is the pulse width $\tau = 10^{-6}/m$.

τ(S)	gain (dB)
10^{-7}	8.0
10^{-8}	13.2
10^{-9}	16.4

It can therefore be seen that the potential gains are quite significant. Note that pulses shorter than one nanosecond do not appear to be feasible (see Chapter 4).

One should bear in mind that the calculations (which are approximate) assume that there is no loss in average power incurred in going from the CW laser to the high peak power system. This is a key assumption and probably overestimates the power advantage of PPM somewhat. Also, assuming that comparable signal-to-noise ratio is required in both systems, ($k_D \approx k_p$) is somewhat favorable to the PPM system. Qualitatively, however, the results indicate that there is considerable potential benefit to be derived in the PPM system, despite the additional complexity of modulation and decoding, if the power trade-off between CW and pulse operation can be made without too much loss in average power.

APPENDIX 10. CANONIC MISSION-MARS ORBITER

This appendix provides the supporting proofs and additional details for the material presented in Chapter 6, Section 1. In this appendix, a series of topics are considered which dictate how certain choices of orbital parameters affect orbital properties important to communication. The results are presented in the simplest meaningful manner, but in each case the extensions necessary to obtain more detailed results will be clear.

1. FRACTION OF TIME A MARS ORBITER IS OCCULTED BY THE PLANET

At typical Earth-Mars distances (say, 1 AU) it is immaterial whether it is assumed that the Mars orbiter is being viewed from an Earth station on the ground, at synchronous altitudes, or even from a libration point tracker. In addition, it may be assumed that the planet casts a cylindrical (rather than conical) shadow at these distances. If a Mars-centered equatorial coordinate system is considered,* and the radius vector to the satellite is expressed as $\vec{r}$ and the unit vector towards the Earth as $\hat{e}$, then the satellite is occulted when

$$\vec{r} \cdot e < -\sqrt{r^2 - R^2} \tag{1}$$

where $r = |\vec{r}|$ and R is the Martian equatorial radius.[1] This is a special case of Equation (16) in this appendix, valid for a conical shadow, which is derived in Section 7 and illustrated in Figure 10-6. Assume that the Earth is located in the x-z plane (this merely provides a reference for the orbiter nodal angle, Ω), at an angle ϵ_M below the equatorial plane, neglecting the difference between the ecliptic and the Mars orbit plane ($\sim 2^\circ$). Finally, assume that the orbiter is in a circular orbit. Then Equation (1) becomes

$$A \cos\theta + B \sin\theta < -\delta \tag{2}$$

where

$$A = \cos\Omega \cos\epsilon_M$$

$$B = -(\sin\Omega \cos i \cos\epsilon_M + \sin i \sin\epsilon_M)$$

$$\delta = \left|h(2R + h)\right|^{1/2}/(R + h). \tag{3}$$

The orbiter altitude and inclination angle have been denoted by h and i respectively, and θ represents the satellite's angular position in the orbit plane measured from the nodal crossing. The angle ϵ_M is variable (from 25 degrees to -25 degrees, over a period of about two years). Computation of ϵ_M as a function of time of year is possible using some simple formulas given in Section 8 where a particular occultation problem is treated from a somewhat different point of view. Here we will simply consider sample results for a specific value of ϵ_M.

The actual entry and exit angles are obtained by replacing the inequality in Equation (2) by an equality. [Also, the angles at which the satellite enters the portion of its orbit directly in front of the planet are given by Equation (2) with + δ on the right hand side.] The fraction of time for which the satellite is occulted is given by

$$f_1 = \psi/\pi \tag{4}$$

where

$$\cos\psi = \delta (A^2 + B^2)^{-1/2} \tag{5}$$

since time and angle are linearly related for circular orbits. A computer program to evaluate f_1 using Equations (3) to (5) with ϵ_M = 25 degrees was used to investigate values of Ω from 0 to 180 degrees, i from 0 to 180 degrees, and h = 0.5R, R, 1.5R, and 2R. Actually, Ω can vary over 360 degrees, but the results for $\Omega > 180$ degrees can be obtained by replacing Ω by Ω + 180 degrees and i by 180 degrees - i in Figures 10-1 to 10-3 where the results are tabulated. [This is to be expected on physical grounds and can also be seen by examination of equation (3).]

Even with all the simplifications that have been made, a four-parameter problem (Ω, i, h, ϵ_M) remains. Actually, the various restrictions (to a circular orbit, cylindrical shadow zone, fixed Earth, spherical Mars, etc.) are easily removed in any particular case,† but so many parameters are then present that it is difficult to present the results in a meaningful manner. Some general statements can be made about Figures 10-1 to 10-3. For given Ω and i values, f_1 increases with decreasing h until $f_1 \to 0.5$ for the limiting case of h = 0. Also, for any Ω or h, there is an i for which f_1 attains a maximum or minimum (or a range of i for which $f_1 = 0$).

*For time intervals of the order of several orbiter periods, this is assumed to be an inertial system.

†See Reference 1. The procedure in general is to obtain entry and exit angles by a numerical iteration procedure, which converges rapidly when started with values obtained from Equation (1).

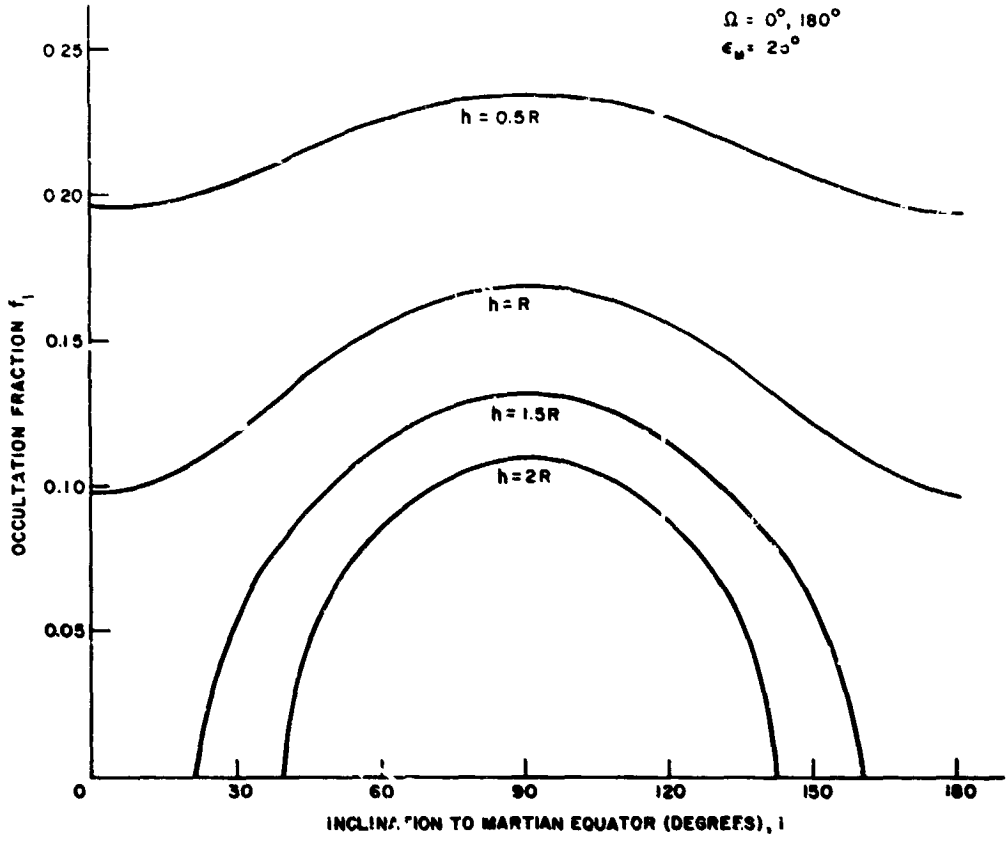


Figure 10-1. f_1 for Ω = 0 degrees, 180 degrees

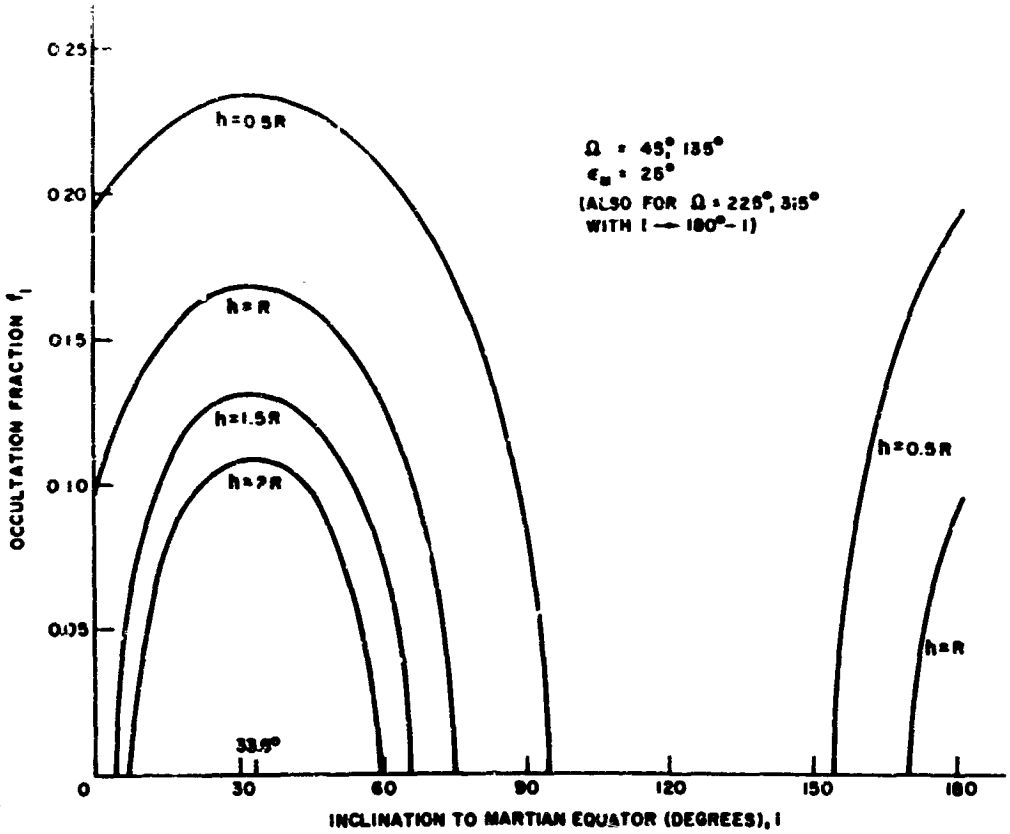


Figure 10-2. f_1 vs. i for Ω = 45 degrees, 135 degrees (and for Ω = 225 degrees, 315 degrees with i → 180 → i)

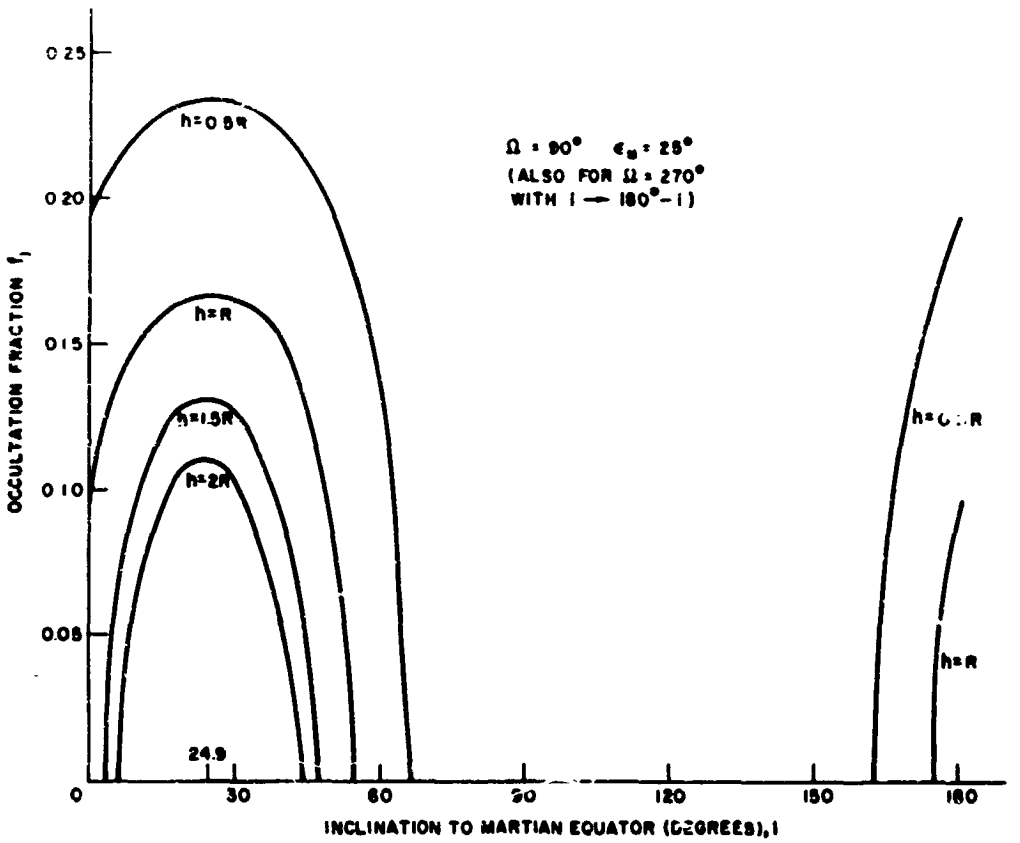


Figure 10-3. f_1 vs. i for Ω = 90 degrees (and for Ω = 270 degrees with i → 180 → i)

Examination of Equations (3) to (5) shows that f_1 is extremized when

$$\tan 2i = 2b/(a-c), \quad \Omega, h \text{ given}$$

where

$$a = \sin^2\Omega\cos^2\epsilon_M$$

$$b = \sin\Omega\cos\epsilon_M \sin\epsilon_M$$

$$c = \sin^2\epsilon_M$$

For $h>0.5R$, it follows that $0<f_1<0.25$. Note that Ω will vary by perhaps 360 degrees per year[2] because of Martian oblateness and also because of the motion of the Earth and Mars about the Sun; therefore, these graphical results are valid only for time intervals of the order of several orbiter periods. Furthermore, since these particular graphs were drawn for $\epsilon_M = 25$ degrees, they are valid for time periods when this is so, although of course graphs could be drawn for any time period by varying ϵ.

In the above discussion, the line of sight to the Mars orbiter is assumed not obscured by the Earth. Loss of visibility may occur for up to 50 percent of the time for trackers on the Earth's surface, up to 5 percent for trackers at synchronous altitudes (see Section 8), or up to 0.6 percent for trackers at the Earth-Moon libration points (see Section 4). It can be avoided by redundant trackers (with attendant handover problems) or by schedules that avoid occultation during periods when communication is particularly important.

2. FRACTION OF TIME MARS IS WITHIN THE BEAM OF THE EARTH TRANSMITTER

To the order of approximation employed in Section 1, the fraction of time that the satellite is in front of the planet is equal to the occultation fraction. Thus,

$$f_2 \approx f_1 + \delta f \tag{6}$$

where δf is a correction due to the finite beamwidth. The additional angular travel during which the beam sees Mars in back of the satellite is approximately

$$2\delta\theta = 2R_E b/(R+h) \tag{7}$$

where $R_E \sim 1$ AU and b is the beamwidth. Thus $R_E b$ is the distance subtended by the beam at Martian distances. For typical values, say $R_E b \approx 500$ miles ($b \approx 0.5\times10^{-4}$ radian $\approx$ 1 arc second) and $h = R$, we have $\delta f = \delta\theta/\pi \approx 0.038$.

Actually, Equation (7) should be used to see the results of variations in b and h, rather than Equation (6) to get exact answers, since the neglect of the conical nature of the shadow in Section 1 is of at least comparable importance to the correction for finite beamwidth (see Section 5).

3. MAGNITUDE AND VARIATION OF DOPPLER SHIFT

The doppler-frequency shift is usually defined as

$$S = -\frac{\nu}{c} V_n \tag{8}$$

where ν is the frequency of radiation emitted from the source, c is the speed of light, and V_n is the component of velocity of the radiation along the line of sight from the source to the receiver. It is simplest here to replace Equation (2) with the equivalent expression

$$S = -\frac{\nu}{c}\frac{d}{dt}|\vec{\rho}| \tag{9}$$

where $\vec{\rho}$ is the vector from source to receiver. Equation (9) must be evaluated for the case in which the source is the Earth (revolving about the Sun) and the receiver is a satellite orbiting around Mars (also revolving about the Sun). Assume that the Mars orbit plane coincides with the ecliptic. Then, if an inertial heliocentric ecliptic system is considered with unit vectors $\hat{x}$, $\hat{y}$, $\hat{z}$ and radial unit vectors are represented by $\hat{r}$, the Sun-Earth vector can be expressed as

$$\vec{r}_E = r_E\hat{r}_E = r_E\left(\hat{x}\cos\theta_E + \hat{y}\sin\theta_E\right) \tag{10}$$

Similarly, for the Sun-Mars vector

$$\vec{r}_M = r_M\hat{r}_M = r_M\left(\hat{x}\cos\theta_M + \hat{y}\sin\theta_M\right) \tag{11}$$

In Equations (10) and (11)

$$\theta_E = n_E t = \frac{2\pi}{P_E} \text{ and } \theta_M = n_M t = \frac{2\pi}{P_M}$$

where the orbital eccentricities of the planets are neglected and time (t) is measured from an Earth-Mars opposition.* Mean motions (angular velocities) and sidereal periods are denoted by n and P with subscripts E and M for Earth and Mars. For simplicity, take $r_E = 1$ AU and $P_E = 1$ year as

*Opposition simply means that the vectors $\hat{r}_E$ and $\hat{r}_M$ are coincident.

units of distance and time. Then $r_M = 1.523679$ and $P_M = 1.8809$. Now we can represent the Earth-Mars vector $\vec{r}_{EM}$ as

$$\vec{r}_{EM} = \vec{r}_M - \vec{r}_E = a\hat{x} + b\hat{y}$$

$$a = r_M \cos\theta_M - \cos\theta_E$$

$$b = r_M \sin\theta_M - \sin\theta_E \qquad (12)$$

Now the vector $\vec{\rho}$ from the Earth to the orbiter is the sum of $\vec{r}_{EM}$ and $\vec{r}$ where $\vec{r}$ represents the orbital motion of the orbiter around Mars. It is convenient to express $\vec{r}$ in the inertial $\hat{x}, \hat{y}, \hat{z}$ system (translated to the Mars-center). Thus

$$\vec{r} = x\hat{x} + y\hat{y} + z\hat{z}$$

$$x = r(\cos\Omega\cos\theta - \sin\Omega\sin\theta\cos i)$$

$$y = r(\sin\Omega\cos\theta + \cos\Omega\sin\theta\cos i)$$

$$z = r\sin\Omega\sin i \qquad (13)$$

For circular orbits, $r = R + h$ and $\theta = nt + \theta_o$ where n is the orbiter mean motion and θ_o is an initial phase angle. Note that the elements in Equation (13) are not the ones discussed in Section 1 in which an Earth-pointing Mars-centered equatorial reference system was considered. Elements in such a rotating reference system cannot be considered constant, even in the absence of perturbations, for time intervals of the length now being considered.

Combining Equations (12) and (13) gives

$$\vec{\rho} = (a+x)\hat{x} + (b+y)\hat{y} + z\hat{z}.$$

For use in Equation (9), it is necessary to compute

$$\frac{d\rho}{dt} = \frac{1}{\rho}\left\{(a+x)\frac{d}{dt}(a+x) + (b+y)\frac{d}{dt}(b+y) + z\frac{dz}{dt}\right\} \qquad (14)$$

Now Equations (9), (12), (13), and (14) can be evaluated numerically for any particular case. Note that the orbital motion of the Mars satellite should make the maximum contribution to S when the orbital plane coincides with the ecliptic. In this case ($i = 0$, $\Omega = 0$), $x = r\cos\theta$, $y = r\sin\theta$, and $z = 0$.

Thus,

$$\frac{d}{dt}(a+x) = 2\pi\left\{-\frac{r_M}{P_M}\sin\theta_M + \sin\theta_E - \frac{r}{P}\sin\theta\right\}$$

$$\frac{d}{dt}(b+y) = 2\pi\left\{\frac{r_M}{P_M}\cos\theta_M - \cos\theta_E + \frac{r}{P}\cos\theta\right\}$$

Insertion of numerical values yields

$$\frac{r_M}{P_M} \sim 0.810$$

$$\frac{r_E}{P_E} = 1.000$$

$$\frac{r}{P} \sim \frac{6200 \text{ miles}}{32400 \text{ seconds}} \sim 0.066 \frac{AU}{year}\ *$$

Thus small errors are made in general by neglecting the orbiter motion. Note also that $a + x \approx a$, $b + y \approx b$, and $\rho \approx (a^2 + b^2)^{1/2} = r_M^2 + 1 - 2r_M\cos(\theta_M - \theta_E)$. Equation (14) produces, after some simplification,

$$\frac{d\rho}{dt} \approx \frac{A\sin\omega t}{1 - B\cos\omega t}$$

$$A = \frac{2\pi r_M(1+P_M)}{P_M\left(1 + r_M^2\right)} \approx 4.44$$

$$B = \frac{2r_M}{1 + r_M^2} \approx 0.917$$

$$\omega = 2\pi\left(\frac{1}{P_M} - 1\right) \approx 2\pi(-0.53) \approx \pi \qquad (15)$$

It is seen that S is approximately periodic with a period of 2 years. Figure 10-4 plots $d\rho/dt$ vs. $\theta = \omega t$. The maximum $d\rho/dt$ occurs at 23.5 degrees (about 1-1/2 months after opposition) and is 11.1 AU/year. Since the units of $d\rho/dt$ are AU/year, they must be divided by $c = 186{,}000$ miles/sec $= 6.31 \times 10^4$ AU/year, to obtain the maximum fractional shift ($1/c\ d\rho/dt$) of 1.76×10^{-4} or about 2.0×10^{-4}. This must be multiplied by ν to obtain the actual value of S.

*That is, for a typical case of $h = 2R$, $r = 3R$.

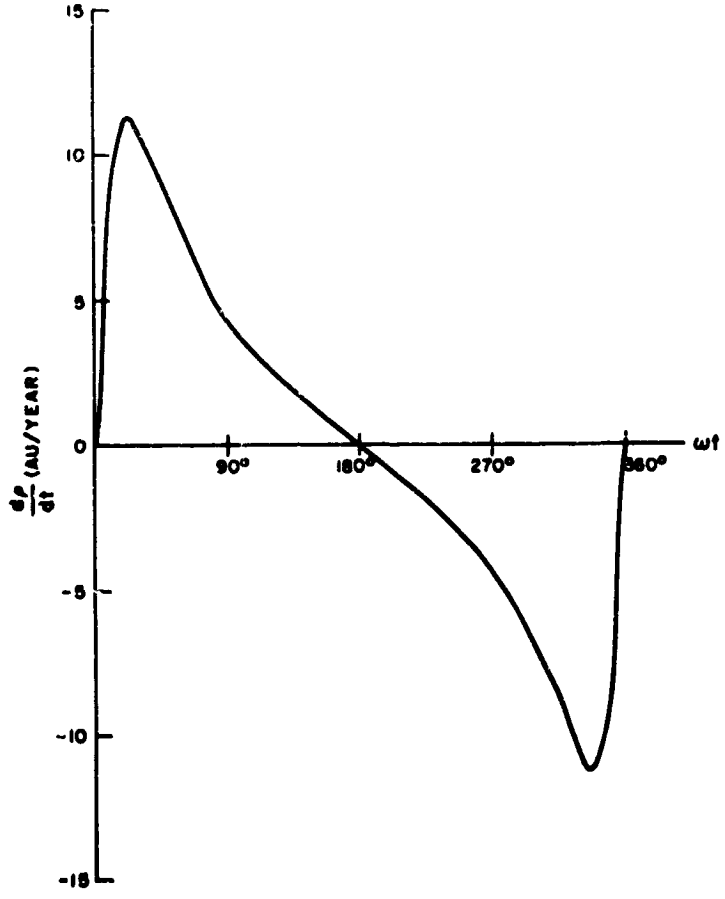


Figure 10-4. Variation of Doppler shift with time

A rough idea of the magnitude of the neglected orbital motion can be obtained by considering a circular orbit in the ecliptic. Then the largest value of V_n [see equation (8)] is just the orbital velocity, which for a typical satellite with h = 2R is about 0.42 AU/year. The small oscillations of this magnitude at most (with period of about 10^{-3} years or $\omega t \approx 0.2$ degrees) should be superimposed on the curve in Figure 10-4.

4. VISIBILITY CONDITIONS FROM A TRACKER SATELLITE SITUATED AT A TRIANGULAR LIBRATION POINT OF THE EARTH-MOON SYSTEM

A complete analysis of this subject would be a rather complex three-dimensional problem, since the Mars orbit plane and the Moon orbit plane are inclined to the ecliptic (by about 2 degrees and 5 degrees, respectively). However, an upper bound can easily be obtained on the fraction of time for which a probe is occulted by considering the problem in the ecliptic plane. The results should be conservative since the slight non-planarities will tend to destroy any colinearity (tracker-Moon-probe or tracker-Earth-probe) that may occur.

Consider Figure 10-5. The Earth and Mars revolve around the Sun with periods of 1 and 1.88 years, respectively. The Moon and the libration point tracker revolve around the Earth during 27.3 days.* A typical Earth-Mars trajectory takes on the order of several hundred days. Thus we can approximate the situation by assuming that the Earth and the probe are approximately fixed relative to each other during one revolution of the Earth-Moon system. Then the fraction of one period (i.e., of 27.3 days) that the probe is occulted by the Moon (see Figure 10-5) is the angle subtended by the Moon at the tracker divided by 360 degrees. (In Figure 10-5 a time of occultation of the probe by the Moon is about to occur.)

The tracker-Earth and tracker-Moon distance is 239,000 miles for a tracker at a triangular libration point, and the Earth and lunar radii are 3970 and 1080 miles. Thus the angles subtended are 1.90 and 0.52 degrees by the Earth and Moon, corresponding to 3.5 and 0.9 hours of occultation per month. This is about 0.6 percent of the time.

Since the angles subtended by the Earth and Moon are comparable or smaller than the skewness of the various orbit planes, which has been neglected in this discussion, one would expect these out-of-plane effects to cause a considerable reduction in our 0.0% maximum occultation fraction. Actually these few hours per month are negligible unless they occur at times when communication is necessary, in which cases they can be avoided by proper scheduling or redundant trackers.

*Actually, the rotation is around the Earth-Moon barycenter which is within the Earth. Thus the order of magnitude arguments are not affected.

5. VISIBILITY PERIODS OF A MARS ORBITER RELATIVE TO A SPACE PROBE APPROACHING FROM EARTH

If space probes approach Mars in the ecliptic plane, then the results of Section 1 (i.e., Figures 10-1 to 10-3) are directly usable, with Ω representing the angle in the ecliptic between the probe-Mars line and the orbiter nodal crossing, as long as the probe is far enough from Mars for our assumption of a cylindrical shadow to be valid. A more general direction of approach could be handled by returning to Equation (1) and letting $\hat{e}$ represent the unit vector towards the probe.

As the probe approaches the planet more closely, the assumption of a cylindrical shadow becomes worse. Eventually, one should replace Equation (1) with

$$\vec{r} \cdot \hat{e} < aR - \sqrt{(1-a^2)(r^2-R^2)} \qquad (16)$$

where $a = R/d$ is the ratio of the Martian radius to the Mars-probe distance. (We derive (16) in Section 7.) For moderate distances Equation (16) can be expanded in a series in a. Then the extended version of Equation (2) becomes

$$A \cos\theta + B \sin\theta < -\delta + \frac{R}{r} a + \frac{1}{2} \delta a^2 + O(a^4) \qquad (17)$$

where $\delta = h(2R + h)^{1/2}/(R + h)$ as before. For example, for h = R, $\delta \approx 0.866$, and the correction terms are negligible until the probe is within a few hundred Mars radii of the planet (d < 0.01 AU). The effect of the approach is to increase f_1, as might be expected, and can be simulated in Figures 10-1 to 10-3 by considering the curves for lower altitude orbiters. Consider as an example a space probe when $a = 0.25$, i.e., at 4 Mars radii from the planet. For an orbiter with h = 2R

$$-\delta + \frac{R}{r} a + \frac{1}{2} \delta a^2 \approx -0.830$$

compared to $-\delta = -0.943$. Since $\delta \approx 0.866$ for h = R, the curves for h = R in Figures 10-1 to 10-3 give approximate values of f_1 for a probe at a distance of about 4R and an orbiter at an altitude of h = 2R. Similar approximations could be made for other distances and altitudes, or Equation (16) could be used directly in each case.

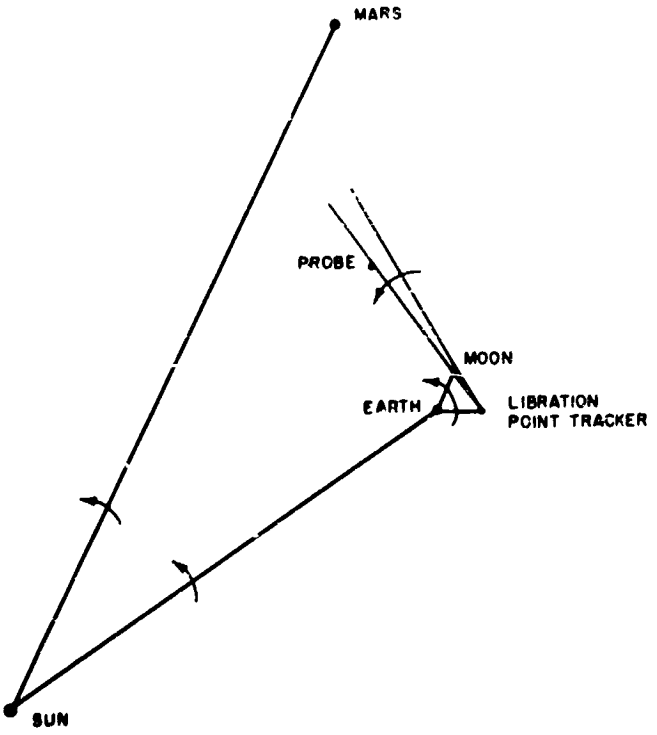


Figure 10-5. Visibility from 'ibration point tracker

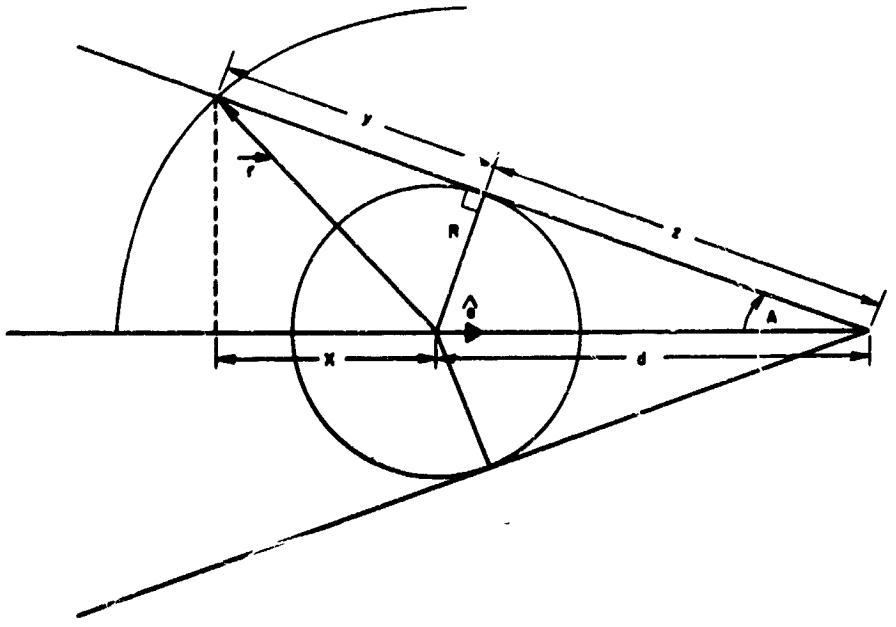


Figure 10-6. Occultation diagram

6. PAYLOAD CONSIDERATIONS FOR A MARS MISSION

The useful payload which can be placed near the target planet is dependent on the type of mission planned and the available booster and vehicle capability, as well as the mission date, and many other factors. Here several of the broad classes of possible missions can be listed and references provided to recent samples of the voluminous literature.

6.1 One-Way Flyby[3]

The probe is launched from Earth or Earth orbit into a heliocentric ellipse past Mars. No primary propulsion is required after Earth departure. This mission's main advantage is its (relative) simplicity as evidenced by the fact that several have already been flown.

6.2 Capture Mission[3]

The probe is made to impact or soft land or orbit Mars by atmospheric or propulsive braking out of the heliocentric transfer ellipse.

6.3 Round-Trip Nonstop Flyby[3]

These are the interplanetary equivalent of lunar free-return trajectories. The features of this mission of interest for communications purposes are close range encounter with the target planet coupled with a return trip during which stored data could be transmitted at relatively low data rates. It is also attractive for an early manned mission. The so-called "powered flyby" in which thrust near Mars is used to augment the effect of the Martian gravity is a variant of this mission. Timing flexibility is gained at the expense of additional complexity.

6.4 Round-Trip Capture or Stopover[3]

These are conceptually simple extensions of the nonstop flyby in which some stopover time on the planet is allowed for in the mission profile.

The above mission classes are basically "ballistic"; that is, only high thrust impulsive propulsion is used during the mission. (In addition, midcourse guidance corrections to reduce errors in the implementation of the impulses would always be required in practice.) Thus minimization of fuel consumption is theoretically equivalent to minimization of instantaneous velocity impulses.

More complex mission types can arise from the above in several ways. One is the allowance of non-impulsive thrust during the mission. For example,[4] continuous low thrust may be used during the entire mission either without impulsive thrust capability or in conjunction with it (mixed or hybrid thrust).

Additional variations are offered by multiplanet flybys or swingby missions. For example,[5] the gravitational field of Venus may be used on the inbound and outbound legs to deflect the probe's trajectory so that acceptable stopover times on Mars are possible at otherwise unfavorable dates. One may consider still more complex combinations such as a powered flyby of Mars (preceded by a Venus swingby) during which a manned excursion vehicle is landed on the Martian surface and then recovered by the main spacecraft.[6] These more complicated missions offer large theoretical fuel savings and/or timing flexibility at the expense of increased complexity.

A relatively simple combination of the basic types, namely a capture mission in which a spacecraft is put in a Martian orbit, followed by a round-trip nonstop flyby offers interesting possibilities of accurate mutual orbit determination between the probe and the orbiter.

7. VISIBILITY CONDITIONS BETWEEN A MARS LANDING VEHICLE AND A MARS ORBITER

Consider the situation in the plane defined by the radius vector to the orbiter ($\vec{r}$) and the unit vector ($\hat{e}$) towards the lending vehicle (LV). From a consideration of Figure 10-6, it can be seen that occultation occurs when

$$\vec{r} \cdot \hat{e} > x \tag{18}$$

that is, when the projection of $\vec{r}$ upon $\hat{e}$ is greater in magnitude than x, and in addition $\vec{r}$ and $\hat{e}$ point towards opposite sides of the occulting body. From Figure 10-6 we see that

$$\begin{aligned} x + d &= (y + z)\cos A \\ y &= (r^2 - R^2)^{1/2} \\ z &= (d^2 - R^2)^{1/2} \\ \cos A &= z/d \end{aligned} \tag{19}$$

Using Equation (19) in Equation (18) we obtain the occultation condition of Equation (16), which we repeat

$$\vec{r} \cdot \hat{e} < aR - \sqrt{(1-a^2)(r^2-R^2)}$$
$$a = R/d \tag{16}$$

[Note that the expression for a cylindrical shadow, (1), follows immediately from Equation (16) for $d \to \infty$, i.e., $a \to 0$.]

The components of the vectors $\vec{r}$ and $\hat{e}$ can be expressed in any convenient coordinate system for the determination of occultation fractions in terms of orbital parameters, as in Section 1.

The situation after the vehicle has landed is described by setting $d = R$ so that $a = 1$ in Equation (16). Thus the condition for occultation becomes

$$\vec{r} \cdot \hat{e} < R \tag{20}$$

where $\hat{e}$ is now the unit vector towards the landed LV. Since the LV may be on the surface for many orbiter periods it is convenient to express Equation (20) in terms of orbital parameters and LV coordinates. Thus, consider that $\vec{r}$ and $\hat{e}$ are expressed in a Mars-centered equatorial coordinate system. Then

$$\begin{aligned} r_1 &= r(\cos\ \Omega \cos\theta - \sin\Omega \sin\theta \cos i) \\ r_2 &= r(\sin\Omega \cos\theta + \cos\Omega \sin\theta \cos i) \\ r_3 &= r \sin\theta \sin i \\ e_1 &= \cos(\alpha+\omega t) \\ e_2 &= \sin(\alpha+\omega t) \\ e_3 &= \sin\delta \end{aligned} \tag{21}$$

where α is the LV longitude in our coordinate system at $t = 0$, δ is the latitude, and ω is the Mars rotation rate. For a circular orbit, $\theta = nt$ and $r = R + h$.

As an example of the use of Equations (20) and (21), consider a satellite at synchronous altitude, $h = h_S$ i.e., such that $n = n_S = \omega$. For an equatorial orbit ($i = 0$), with the satellite at $t = 0$ above the LV ($\Omega = \alpha = \delta = 0$), Equation (20) becomes

$$\cos\omega t \cdot (R+h_S)\cos n_S t + \sin\omega t \cdot (R+h_S)\sin n_S t < R$$

or

$$\cos(\omega - n_S)t < \frac{R}{R+h_S}$$

for occultation. Since $\omega = n_S$ and $h_S > 0$, the LV is always visible from the satellite, as of course it should be in this case. More complex cases may be treated directly from Equations (20) and (21), and occultation fractions may be found as in Section 1.

8. VISIBILITY OF A MARS SYNCHRONOUS SATELLITE FROM AN EARTH SYNCHRONOUS SATELLITE

It has been shown previously that under certain conditions a Mars orbiter would not be occulted by Mars for at least several orbiter periods, and thus will remain visible from the Earth if occultation by Mars alone is considered. Of course, in practice occultation by the Earth must also be considered. In Section 1, an easily obtained upper bound, was noted, namely that occultation by the Earth when viewing from a Earth synchronous satellite occurs at most 5 percent of any given day.

It is of interest to know whether, as one might expect, there are extended periods when no occultation occurs and continuous communication is possible. Although purely geometrical, the problem is somewhat involved and the proverbial slide rule estimates cannot be relied upon. The situation is analyzed in the following sections.

8.1 Visibility of Mars from an Earth Synchronous Satellite

Consider a heliocentric inertial system with axis x' pointing towards the first point of Ares (γ), z' normal to the ecliptic (directed north), and y' in the ecliptic such that x', y', and z' form a right-handed system. (See Figure 10-7.) Thus the x' - y' plane is the ecliptic plane. The Mars orbit plane intersects the Earth orbit plane in an axis x, an angle Ω_M from γ. The orbit planes are inclined at an angle i_M. Denote the angular distance of Mars from x in the Mars orbit plane by θ_M and the angular distance of the Earth from x in the ecliptic by θ_E. Assume

$$\theta_M = \frac{n_M}{n_E}\theta_E \tag{22}$$

where n_M and n_E are the Martian and terrestrial mean motions. Equation (22) assumes time is measured from an Earth-Mars opposition, i.e., when both are on the x axis. These occur approximately every 2 years. Denote the Sun-Earth and Sun-Mars distances by ρ_E and ρ_M, respectively, and assume them constant. (The quantities defined so far are illustrated in Figure 10-7, the Earth, Mars, and the Sun being denoted by E, M, and S respectively.) At any moment the Earth's equatorial plane is inclined below the ecliptic an angle e.

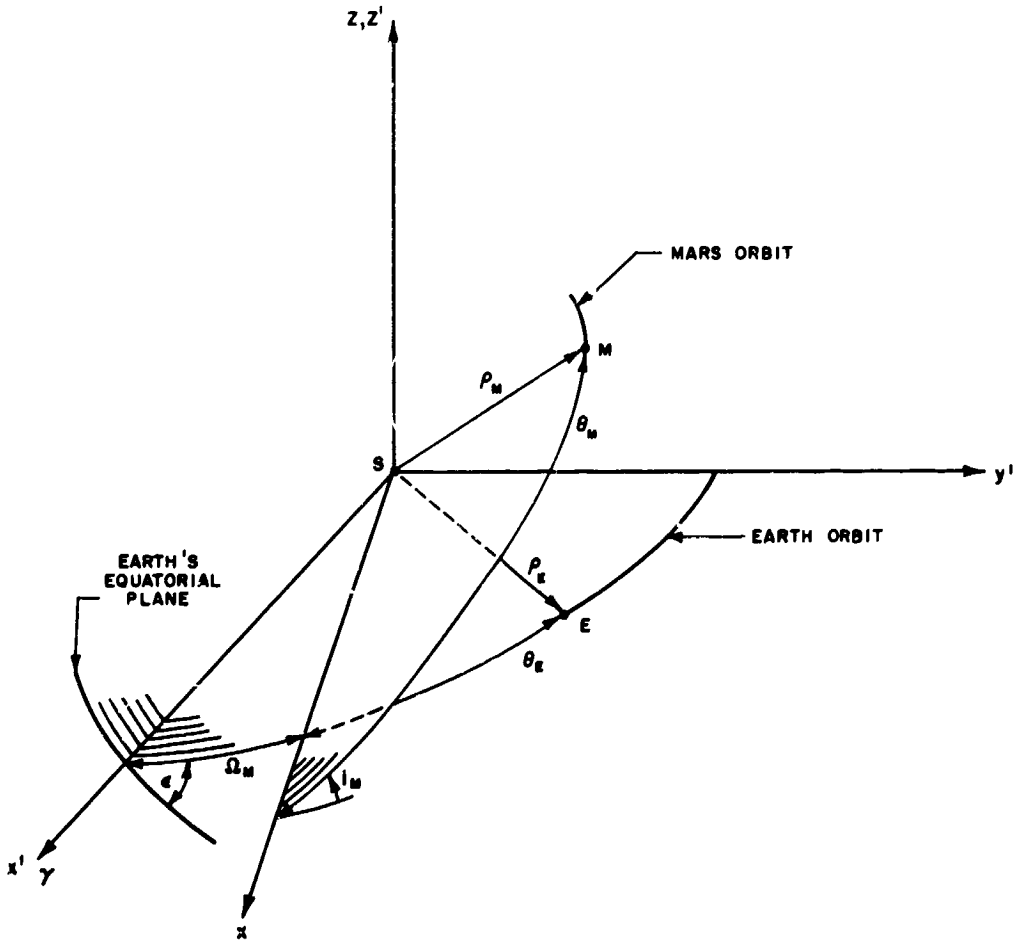


Figure 10-7. Orbital geometry

The coordinates must be obtained of the Earth-Mars distance in a coordinate system x", y", and z" where x" and y" lie in the Earth's equatorial plane and z" points to the North Pole. First note that

$$x = \frac{\rho_M}{\rho_E}\cos\theta_M - \cos\theta_E$$

$$y = \frac{\rho_M}{\rho_E}\sin\theta_M\cos i_M - \sin\theta_E$$

$$z = \frac{\rho_M}{\rho_E}\sin\theta_M\sin i_M \qquad (23)$$

where x, y, and z are the components of the Earth-Mars distance in astronomical units. The x', y', and z' system is obtained from x, y, and z by a counterclockwise rotation of Ω_M around z. Then x", y", and z" are obtained from x', y', and z' by a clockwise rotation of ϵ around x'. Thus

$$x'' = x\cos\Omega_M - y\sin\Omega_M$$

$$y'' = \cos\epsilon(x\sin\Omega_M + y\cos\Omega_M) - z\sin\epsilon$$

$$z'' = \sin\epsilon(x\sin\Omega_M + y\cos\Omega_M) + z\cos\epsilon \qquad (24)$$

and x", y", and z" can be considered to be the coordinates of Mars in an Earth-centered equatorial coordinate system.

Now consider an Earth-synchronous equatorial satellite S_1 in Figure 10-8. S_1 is a distance r_s from the Earth which subtends an angle $2\phi_s$ at S_1. The coordinates of Mars $(x''^2 + y''^2)^{1/2}$, z" are approximately constant during one day; during this time the shadow cone (vertex at S_1, vertex angle $2\phi_s$) sweeps out the solid of revolution whose cross section is shown in Figure 10-8. For continuous visibility M must be outside this region, or

$$|z''| > \left\{ r_s + (x''^2 + y''^2)^{1/2} \right\} \tan\phi_s \qquad (25)$$

Equation (25) is an inequality dependent upon θ_E through Equations (22) to (24). A computer program to evaluate Equations (22) to (25) was prepared and run with the following constants.

$$\frac{\rho_M}{\rho_E} = 1.524$$

$$\frac{n_M}{n_E} = 0.531$$

$$i_M = 1.85^\circ$$

$$\Omega_M = 49.33^\circ \quad \text{(epoch 1971)}$$

$$\epsilon = 23.45^\circ$$

$$r_s = 0.26 \times 10^{-3} \text{ AU}$$

$$\phi_s = 8.75^\circ$$

The results as a function of θ_E are*

$0^\circ \leq \theta_E \leq 252^\circ$: visible for 252°

$253^\circ \leq \theta_E \leq 313^\circ$: not visible for 60°†

$314^\circ \leq \theta_E \leq 506^\circ$: visible for 192°

$507^\circ \leq \theta_E \leq 571^\circ$: not visible for 64°

$572^\circ \leq \theta_E \leq 950^\circ$: visible for 378°

$951^\circ \leq \theta_E \leq 1013^\circ$: not visible for 62°

etc.

8.2 Occultation of a Mars Orbiter by Mars

Consider a Mars-centered inertial system with the Z axis pointing north along the Mars axis of rotation, the Y axis along the intersection of the ecliptic and the Mars equatorial plane, and the X axis completing the right-handed system (see Figure 10-9). The ecliptic intersects the X-Z plane in a line an angle ϵ_M below the X axis where $\epsilon_M \approx 25$ degrees. (The inclination of the Mars orbit plane to the ecliptic, $i_M = 1.85$ degrees, will be neglected in the calculations.) A unit vector $\hat{e}$ pointing in the Mars-Earth direction makes an angle θ_E with the X-Z plane.

Assume that Mars casts a cylindrical shadow; thus the satellite is occulted when

$$\vec{r} \cdot \hat{e} < \sqrt{r^2 - R} \qquad (26)$$

*A change in θ_E of 1 degree corresponds to 365.25 .../360 days, or about 1 day.

†"Not visible" means not continuously visible.

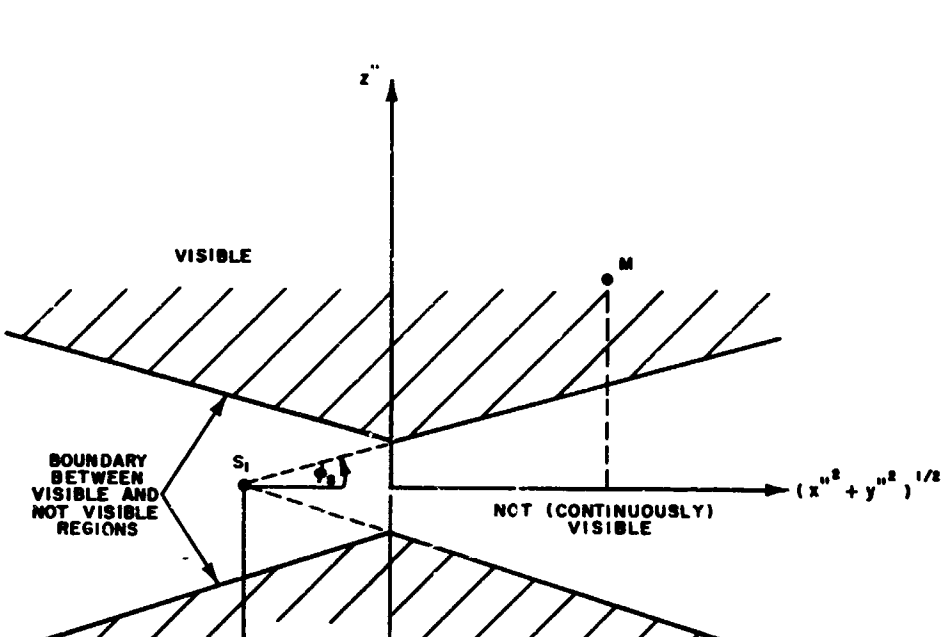


Figure 10-8. Visibility geometry

where $\vec{r}$ is the radius vector from Mars to the satellite, $r = |\vec{r}|$, and R is the Mars equatorial radius. The vector $\hat{e}$ has components

$$\begin{aligned} e_x &= \cos \epsilon_M \cos \theta_E \\ e_y &= \sin \theta_E \\ e_z &= -\sin \epsilon_M \cos \theta_M \end{aligned} \tag{27}$$

in the system M-XYZ. Similarly the vector $\vec{r}$ is expressed as

$$\begin{aligned} r_x &= r(\cos \Omega \cos \theta - \sin \Omega \cos i \sin \theta) \\ r_y &= r(\sin \Omega \cos \theta + \cos \Omega \cos i \sin \theta) \\ r_z &= r \sin i \sin \theta \end{aligned} \tag{28}$$

where the angles Ω, θ, and i specify the direction of $\vec{r}$ as shown in Figure 10-9. Using Equations (27) and (28) in Equation (26), the occultation condition is*

$$A \cos \theta + B \sin \theta < -\delta \tag{29}$$

where

$$\begin{aligned} A &= \cos \epsilon_M \cos \bar{\theta}_E \cos \Omega + \sin \bar{\theta}_E \sin \Omega \\ B &= -\cos \epsilon_M \cos \bar{\theta}_E \sin \Omega \cos i + \sin \bar{\theta}_E \cos \Omega \cos i \\ &\quad - \sin \epsilon_M \cos \bar{\theta}_E \sin i \\ \delta &= \frac{1}{R+h}\left[h(2R+h)\right]^{1/2} \end{aligned} \tag{30}$$

The altitude h in Equation (30), given by $h = r - R$ is a constant if the orbiter is in a circular orbit, as assumed below.

*Equations (29) and (30) correspond to Equations (2) and (3) of Section 1; however, they are equivalent only when $\bar{\theta}_E = 0$. Thus the graphical results for f_1 obtained in Section 1 are valid only for periods of time around $\bar{\theta}_E = 0$. Results for other time periods could be tabulated graphically by replacing ϵ_M in Section 1 by an angle ϵ' such that

$$\sin \epsilon' = \sin \epsilon_M \cos \bar{\theta}_E$$

The actual angles, θ, for shadow-zone entry and exit are given by replacing the inequality in Equation (29) by an equality. The occultation fraction is given by

$$f_1 = \frac{\psi}{\pi} \tag{31}$$

where

$$\cos \psi = \delta \left(A^2+B^2\right)^{-1/2} \tag{32}$$

For the case of equatorial Mars orbiters, using i = 0 in Equation (30),

$$A^2 + B^2 = \sin^2 \bar{\theta}_E + \cos^2 \epsilon_M \cos^2 \bar{\theta}_E$$

Thus occultation <u>never</u> occurs when

$$1 - \left(\frac{R}{r}\right)^2 > \sin^2 \bar{\theta}_E + \cos^2 \epsilon_M \cos^2 \bar{\theta}_E$$

or when

$$|\cos \bar{\theta}_E| > \{R\,(R+h) \sin \epsilon_M\}^{-1} \tag{33}$$

Note that there is no h for which occultation never occurs. However, the period of the year (i.e., region of $\bar{\theta}_E$) can be found for which there is no occultation for a given altitude satellite.

8.3 Continuous Visibility

For continuous visibility, Equations (25) and (33) must be satisfied simultaneously so that the Earth synchronous satellite can see Mars without being occulted by the Earth, and the Mars synchronous satellite is visible from the Earth without being occulted by Mars. To check the simultaneous satisfaction of Equations (25) and (33), $\bar{\theta}_E$ as a function of θ_E must be known.

Referring to Figure 10-10, the orientation of the normal to the Mars equatorial plane, line SN, is given by right ascension and declination angles, α_o and δ_o, with respect to Aries and the Earth's equator where[7]

$$\begin{aligned} \alpha_o &\approx 317.93^\circ \\ \delta_o &\approx 54.73^\circ \end{aligned} \qquad \text{(epoch 1971)}$$

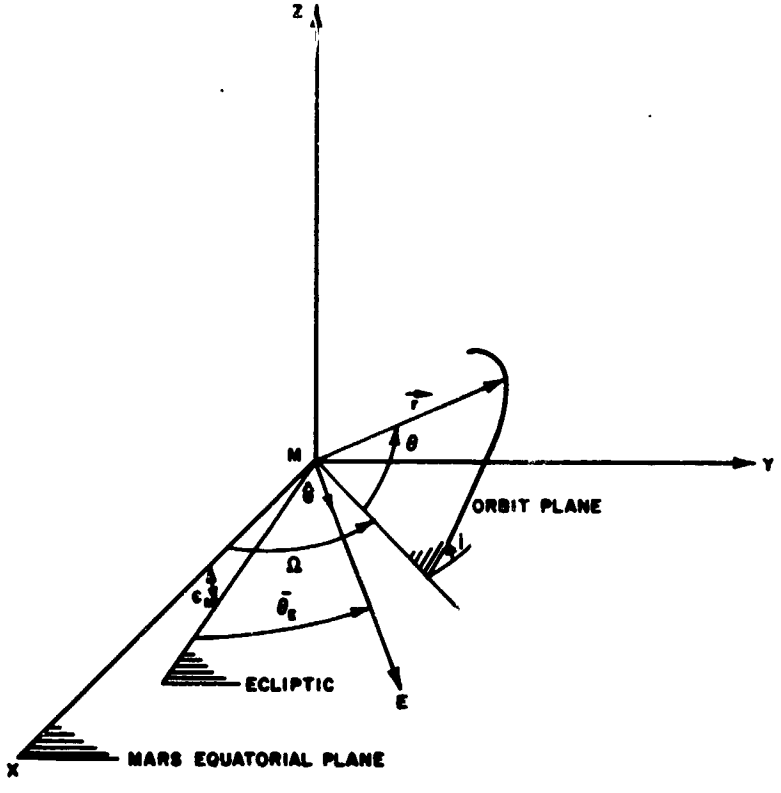


Figure 10-9. Mars orbiter geom try

The orientation of the line of intersection of the Mars equatorial plane, SN' and the ecliptic can be defined by an angle β, measured in the ecliptic from γ.

The direction cosines of SN are

$$\left| \cos\delta_o \cos\alpha_o, \cos\delta_o \sin\alpha_o, \sin\delta_o \right|$$

those of SN' are

$$\left| \cos\beta, \sin\beta\cos\epsilon, \sin\beta\sin\epsilon \right|$$

The orthogonality of SN and SN' can be used to obtain an expression for tan β, namely,

$$\tan\beta = -\frac{\cos\alpha_o}{\cos\epsilon \sin\alpha_o + \sin\epsilon \tan\delta_o}$$

yielding $\beta = 85.98^{\circ}$.*

Now consider the following directions and angles in the ecliptic (see Figure 10-11). Let

γ indicate the direction of the line of intersection of the Earth's equatorial plane and the ecliptic.

A indicate the direction of the line of intersection of the Mars equatorial plane and the ecliptic (i.e., a direction parallel to SN' in Figure 10-10.)

$\alpha = \bar{\theta}_E - 90$ degrees be the angle between A and the M-E line

γ be the angle between γ and A

$\gamma = \Omega_E + \theta_E$ be the angle between γ and the S-E line

$\delta_M = \Omega_M + \theta_M$

The distances S-E and S-M are 1 and ρ_M/ρ_E in AUs. Thus the components of the distance $|\vec{ME}|$ are given by

$$|\vec{ME}|_\gamma = \cos\delta - \frac{\rho_M}{\rho_E}\cos\delta_M = |\vec{ME}|\cos\phi$$

$$|\vec{ME}|_n = \sin\delta - \frac{\rho_M}{\rho_E}\sin\delta_M = |\vec{ME}|\sin\phi \qquad (34)$$

where $\phi = \alpha + \beta$ is the angle between γ and the M-E line. Thus, since

$$\phi = \alpha + \beta = \bar{\theta}_E + \beta - 90^{\circ}$$

then

$$\bar{\theta}_E = \phi - (\beta + 90^{\circ}) \qquad (35)$$

where ϕ is given in terms of θ_E using Equation (34).

A computer program to evaluate Equation (33), with θ_E as the independent variable, has been written. It is found that

$0^{\circ} \leq \theta_E \leq 102^{\circ}$: no occultation for 102°

$103^{\circ} \leq \theta_E \leq 178^{\circ}$: occultation for 75°

$179^{\circ} \leq \theta_E \leq 369^{\circ}$: no occultation for 190°

$370^{\circ} \leq \theta_E \leq 434^{\circ}$: occultation for 64°

$435^{\circ} \leq \theta_E \leq 628^{\circ}$: no occultation for 193°

$529^{\circ} \leq \theta_E \leq 867^{\circ}$: occultation for 239°

$868^{\circ} \leq \theta_E \leq 1070^{\circ}$: no occultation for 202°

etc.

Combining these with the results found in Section 8.1 shows there is continuous visibility when

$0^{\circ} \leq \theta_E \leq 102^{\circ}$ (102°)

$179^{\circ} \leq \theta_E \leq 252^{\circ}$ (73°)

$314^{\circ} \leq \theta_E \leq 369^{\circ}$ (55°)

$435^{\circ} \leq \theta_E \leq 506^{\circ}$ (71°)

$572^{\circ} \leq \theta_E \leq 628^{\circ}$ (56°)

$868^{\circ} \leq \theta_E \leq 950^{\circ}$ (82°)

$1014^{\circ} \leq \theta_E \leq 1070^{\circ}$ (56°)

etc.

Thus there are periods of 2 or 3 months, 2 or 3 times per year when continuous communication is possible between an Earth synchronous satellite and a Mars synchronous satellite.

*There is an ambiguity of ±180 degrees in β. However, this does not affect the subsequent calculations, as may be seen by replacing β by $\beta = \beta \pm 180$ degrees. (Note that α then = α – 180 degrees.)

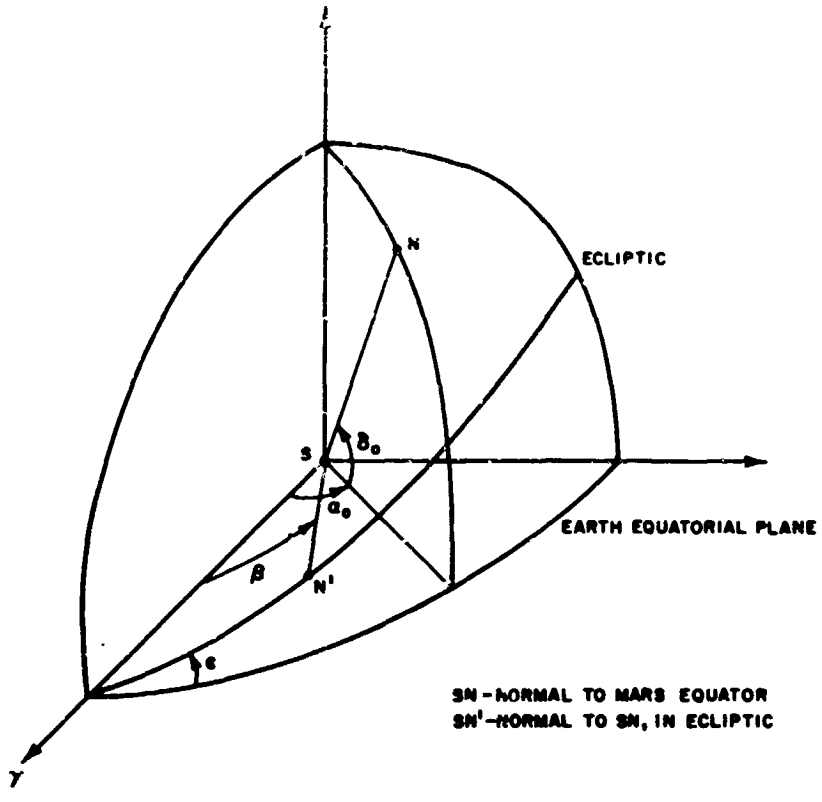


Figure 10-10. Orientation of intersection of mars equatorial plane with ecliptic

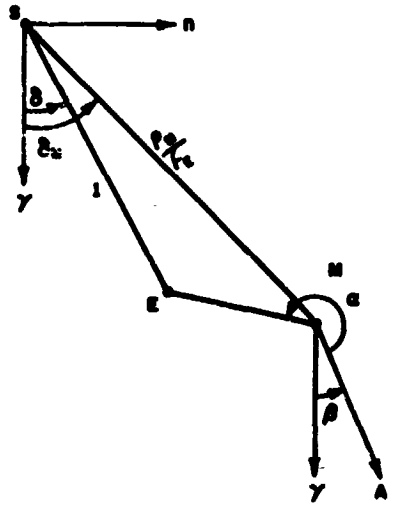


Figure 10-11. Ecliptic plane geometry

REFERENCES

1. P.R. Escobal, "Orbital Entrance and Exit from the Shadow of the Earth," ARS Journal, 32, No. 12 (December 1962), pp 1939-1942.

2. R.L. Moll and M.A. Krop, "Long Lifetime Orbits About Mars," AIAA Paper No. 66-35, presented at the 3rd Aerospace Sciences Meeting, 1966.

3. V.A. Lee and S.W. Wilson, Jr., "A Survey of Ballistic Mars-Mission Profiles," Journal of Spacecraft and Rockets (February 1967), pp 129-142 (with 53 references).

4. R.V. Ragsac "Study of Electric Propulsion for Manned Mars Missions," Journal of Spacecraft and Rockets (April 1967), pp 462-468.

5. R.W. Gillespie and S. Ross, "Venus-Swingby Mode and its Role in the Manned Exploration of Mars," Journal of Spacecraft and Rockets (February 1967), pp 170-175.

6. R.L. Solin, "Mars/Venus Flyby Missions with Manned Mars Landers," Journal of Spacecraft and Rockets (January 1967), pp 115-117.

7. Explanatory Supplement to the Ephemeris, Her Majesty's Stationary Office, London, 1961, p 336.

APPENDIX II.

SENSITIVITIES OF HYPERBOLIC AND ELLIPTIC ORBITS

1. The Hyperbolic Orbit

The position vector of the hyperbolic probe orbit is taken to be

$$\underline{\rho} = \hat{1}\rho\cos\theta + \hat{2}\rho\sin\theta$$

where $\rho = a_H[e\cosh F - 1]$ and $\theta = f + \omega$. The following relations hold among the elements of the orbit.

a_H = semimajor axis

$n = k^{1/2}a_H^{-3/2}$ = mean motion

ω = longitude of pericenter

f = true anomaly

F = hyperbolic anomaly

τ_H = time of pericenter passage

$t = \tau_H + [e\sinh F - F]/n$

$$F = 2\tanh^{-1}x = \ln\left(\frac{1+x}{1-x}\right)$$

$$x = \left(\frac{e-1}{e+1}\right)^{1/2}\tan f$$

$$\frac{dt}{df} = \frac{\rho^2}{na^2(e^2-1)^{1/2}}$$

Given the elements a_H, e, τ_H, and ω, the partial derivatives of the rectangular coordinates and velocities can be found with respect to the elements. Let $\psi = (e^2-1)^{1/2}$, $\rho = a_H\psi^2(1 + e\cos f)^{-1}$, $m = e\sin\omega$, $\ell = e\cos\omega$.

$$\frac{\partial\rho_1}{\partial a_H} = \frac{\rho}{a_H}\cos\theta + \frac{3n}{2\dot{\psi}}(t-\tau)(m+\sin\theta)$$

$$\frac{\partial\rho_1}{\partial e} = a_H\cos\omega$$

$$\frac{\partial\rho_1}{\partial\omega} = -\rho_2$$

$$\frac{\partial\rho_1}{\partial\tau_H} = \frac{na_H}{\dot{\psi}}(m+\sin\theta)$$

$$\frac{\partial\rho_2}{\partial a_H} = \frac{\rho}{a_H}\sin\theta - \frac{3n}{2\dot{\psi}}(t-\tau)(\ell+\cos\theta)$$

$$\frac{\partial\rho_2}{\partial e} = a_H\sin\omega$$

$$\frac{\partial\rho_2}{\partial\omega} = \rho_1$$

$$\frac{\partial\rho_2}{\partial\tau_H} = -\frac{na_H}{\dot{\psi}}(\ell+\cos\theta)$$

$$\frac{\partial\dot{\rho}_1}{\partial a_H} = \frac{n(m+\sin\theta)}{2\dot{\psi}} + \frac{3n^2a_H^2(t-\tau)\cos\theta}{2\rho^2}$$

$$\frac{\partial\dot{\rho}_1}{\partial e} = 0$$

$$\frac{\partial\dot{\rho}_1}{\partial\omega} = -\frac{na_H}{\dot{\psi}}(\ell+\cos\theta) = -\dot{\rho}_2$$

$$\frac{\partial\dot{\rho}_1}{\partial\tau_H} = \frac{k\cos\theta}{\rho^2}$$

$$\frac{\partial\dot{\rho}_2}{\partial a_H} = -\frac{n}{2\dot{\psi}}(\ell+\cos\theta) + \frac{3n^2a_H^2(t-\tau)}{2\rho^2}\sin\theta$$

$$\frac{\partial\dot{\rho}_2}{\partial e} = 0$$

$$\frac{\partial\dot{\rho}_2}{\partial\omega} = -\frac{na_H}{\dot{\psi}}(m+\sin\theta) = \dot{\rho}_1$$

$$\frac{\partial\dot{\rho}_2}{\partial\tau_H} = \frac{k\sin\theta}{\rho^2}$$

2. The Elliptic Orbit

Elements used for the elliptic (satellite) orbit are

a_E = semimajor axis

$\ell = e\cos\omega$

$m = e\sin\omega$

τ_E = time of pericenter passage.

Solution of the standard Kepler equation to determine position in the elliptic orbit and partial derivatives of the latter with respect to the elements was available from earlier tracking studies in terms of a library routine.

APPENDIX 12.
HETERODYNE DETECTION OF OPTICAL SIGNALS WITH A PHASED ARRAY

As has already been shown,[1,2,3] if the collecting aperture of a heterodyne optical receiving system exceeds the coherence area of the signal field, then the signal-to-noise ratio (SNR) saturates at a level determined by this coherence area. Rather than considering a single large collecting area, it is then natural to consider a phased array of smaller apertures, each with the area no larger than the coherence area of the signal field. This is illustrated schematically in Figure 12-1, where signal and local oscillator are incident on N apertures. In each of these N channels, the resultant field is photodetected and the output current is filtered at the difference frequency. The output of each filter consists of a signal term, $s_i(t)$, which is proportional to the product of the signal and local oscillator fields, and a shot noise term, $n_i(t)$, whose standard deviation is proportional to the magnitude of the local oscillator field.* Because the shot noise is observed in a narrow bandwidth W, it is reasonable to treat the $n_i(t)$ as Gaussian processes. It will be assumed further that the $n_i(t)$ are independent processes and are independent of the signal.

The complex envelope representation[4] of narrowband signals will be employed. In this representation, n_i (evaluated at a fixed instant of time†) is a radially symmetric complex Gaussian variate with‡ $En_i = 0 = En_i^2$, $E|n_i|^2 = 2\sigma^2$. The signals $s_i(t)$ are assumed to be of the form

$$s_i(t) = A \exp[j\theta_i(t)] \tag{1}$$

where the θ_i are independent random variables uniformly distributed on $(0,2\pi)$.

The outputs of the N channels are combined (in manners to be discussed below) to give a resultant output R(t). The output SNR will be defined as

$$SNR \equiv \frac{\sqrt{2}\,(E_s|R|^2 - E_n|R|^2)}{(\mathrm{var}_s|R|^2 + \mathrm{var}_n|R|^2)^{1/2}} \tag{2}$$

where the subscript s denotes the case where signal and noise are present, and the subscript n the case where noise alone is present. It is generally more common[1,2,3] to define SNR by the ratio of the expectation of $|R|^2$ when signal alone is present, to the expectation of $|R|^2$ when noise alone is present. However, in the case of digital communication systems where we wish to decide between the two hypotheses: either signal plus noise or noise alone is present, then the second is the more appropriate performance criterion. Indeed, it has been shown that the more conventional definition can in some cases lead to grossly misleading results.

Three methods of combining will be considered: (1) linear, (2) square law, and (3) linear with phase correction (matched filter); and these will be compared on the basis of the input SNR, $A^2/2\sigma^2$, required to achieve a given SNR.

1. LINEAR COMBINING

Consider

$$R_l = \sum_{i=1}^{N} (Ae^{j\theta_i} + n_i) \tag{3}$$

In this case, it can be shown that

$$E_s|R_l|^2 = NA^2 + 2N\sigma^2$$
$$\mathrm{var}_s|R_l|^2 = (N^2 - N)A^4 + 4N^2A^2\sigma^2 + 4N^2\sigma^4$$

and the corresponding moments for the subscript n are obtained by replacing A by zero. Therefore,

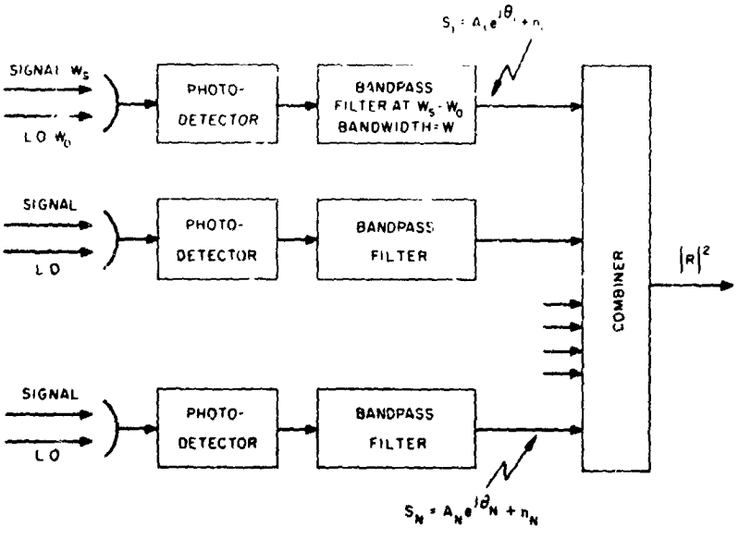
$$(SNR)_l = \frac{A^2/2\sigma^2}{\sqrt{1 + \dfrac{A^2}{2\sigma^2} + \left(1 - \dfrac{1}{N}\right)\dfrac{A^4}{8\sigma^2}}} \tag{4}$$

Note that for large N the SNR is essentially independent of N. Note also that, for $N > 1$, the output SNR also saturates with increasing A/σ, so that improvement is not achieved by increasing the input signal-to-noise ratio. The

*It is assumed that the local oscillator is much stronger than the signal. The shot noise variance is proportional to the DC current output of the photodetector, which (with the above assumption) is proportional to the magnitude squared of the local field.

†When we omit the argument of the random process, we will always be considering the random variable corresponding to a fixed instant of time.

‡E denotes expectation.



1 LINEAR COMBINER $|R|^2 = \left|\sum A_i e^{j\theta_i} + n_i\right|^2$

2 SQUARE-LAW COMBINER $|R|^2 = \sum \left|A_i e^{j\theta_i} + n_i\right|^2$

3. COHERENT COMBINER $|R|^2 = \left|\sum A_i + n_i e^{-j\theta_i}\right|^2$

Figure 12-1. Phased array heterodyne

reason for this is that random phase addition gives essentially a Gaussian signal for which the average power and the standard deviation of the average power are proportional to one another.

Thus linear combining affords no advantage, and indeed can result in performance considerably poorer than a single channel. Note that linear combining is essentially what occurs when a collecting aperture larger than the coherence area is employed.

2. SQUARE-LAW COMBINING

Consider

$$R_2 = \left\{ \sum_{i=1}^{N} |Ae^{j\theta_i} + n_i|^2 \right\}^{1/2} \qquad (5)$$

in which case

$$E_s |R_2|^2 = NA^2 + 2N\sigma^2$$

$$\mathrm{var}_s |R_2|^2 = 4NA^2\sigma^2 + 4N\sigma^4$$

Therefore

$$(SNR)_2 = \frac{N A^2/2\sigma^2}{\sqrt{1 + A^2/2\sigma^2}} \qquad (6)$$

It follows from Equation (6) that to achieve a fixed $(SNR)_2$ the input signal-to-noise ratio $(A^2/2\sigma^2)$ may be decreased proportional to $1/N$ when $A^2/2\sigma^2 \gg 1$, but decreases proportional to $1/\sqrt{N}$ when $A^2/2\sigma^2 \ll 1$. The latter case is the well-known incoherent integration result.

3. COHERENT COMBINING

If the individual θ_i were known, phase correction could be employed to obtain

$$R_3 = \sum_{i=1}^{N} \left(A + n_i e^{-j\theta_i} \right) \qquad (7)$$

in which case

$$E_s |R_3|^2 = N^2A^2 + 2N\sigma^2$$

$$\mathrm{var}_s |R_3|^2 = 4N^3A^2\sigma^2 + 4N^2\sigma^4$$

which gives

$$(SNR)_3 = \frac{NA^2/2\sigma^2}{\sqrt{1 + \frac{NA^2}{2\sigma^2}}} \qquad (8)$$

Here the input signal-to-noise ratio required to achieve a fixed output $(SNR)_3$ decreases as $1/N$, corresponding to the well-known coherent integration result.

4. DISCUSSION

The input signal-to-noise ratio $(A^2/2\sigma^2)$ required to achieve SNR = 4* is shown in Figure 12-2 as a function of N for square-law combining [Equation (6)] and coherent combining [Equation (8)]. Clearly, the results are identical for N = 1. Coherent combining (matched filter) permits a 3dB reduction in $A^2/2\sigma^2$ for each doubling of N, whereas in the limit of large N, square-law combining permits only a 1.5 dB reduction. However, for moderate values of N the performance of square-law combining is not much poorer than the optimum. For example, for N = 10, there is only a 1.3 dB difference between the two curves.

To perform the optimum coherent combining it is necessary to know the phases θ_i. However, there is generally inadequate signal to noise to measure these phases over the full communication bandwidth W.* If the phase variation is, however, independent of frequency, then it is possible to use a narrowband B($B \ll W$) component of the signal (e.g., a carrier or pilot as in the STAR repeater[5]) to obtain the phase. The ratio of carrier power P_c to signal power P_s required for this is given approximately by

$$\frac{P_c}{P_s} \approx \frac{NB}{W} \qquad (9)$$

The basis of Equation (9) is that, if coherent combining is required, the signal power per channel is a factor N too small for satisfactory detection. Thus, we require this factor more carrier power, reduced of course by the bandwidth ratio B/W. It is desirable to make B as small as possible, the lower limit being determined by the bandwidth of the phase fluctuation. Typically this will be less than 1 KHz so that, for a 1 MHz communication bandwidth, 100 channels could be combined utilizing a carrier power of the order of 10 percent of the communication power.

*Note that, from Equation (2), this corresponds to requiring that, for a binary error to be made, the "test statistic" differ by four standard deviations from its mean.

*If there were sufficient signal-to-noise ratio in each channel, then combining improvement would not be required.

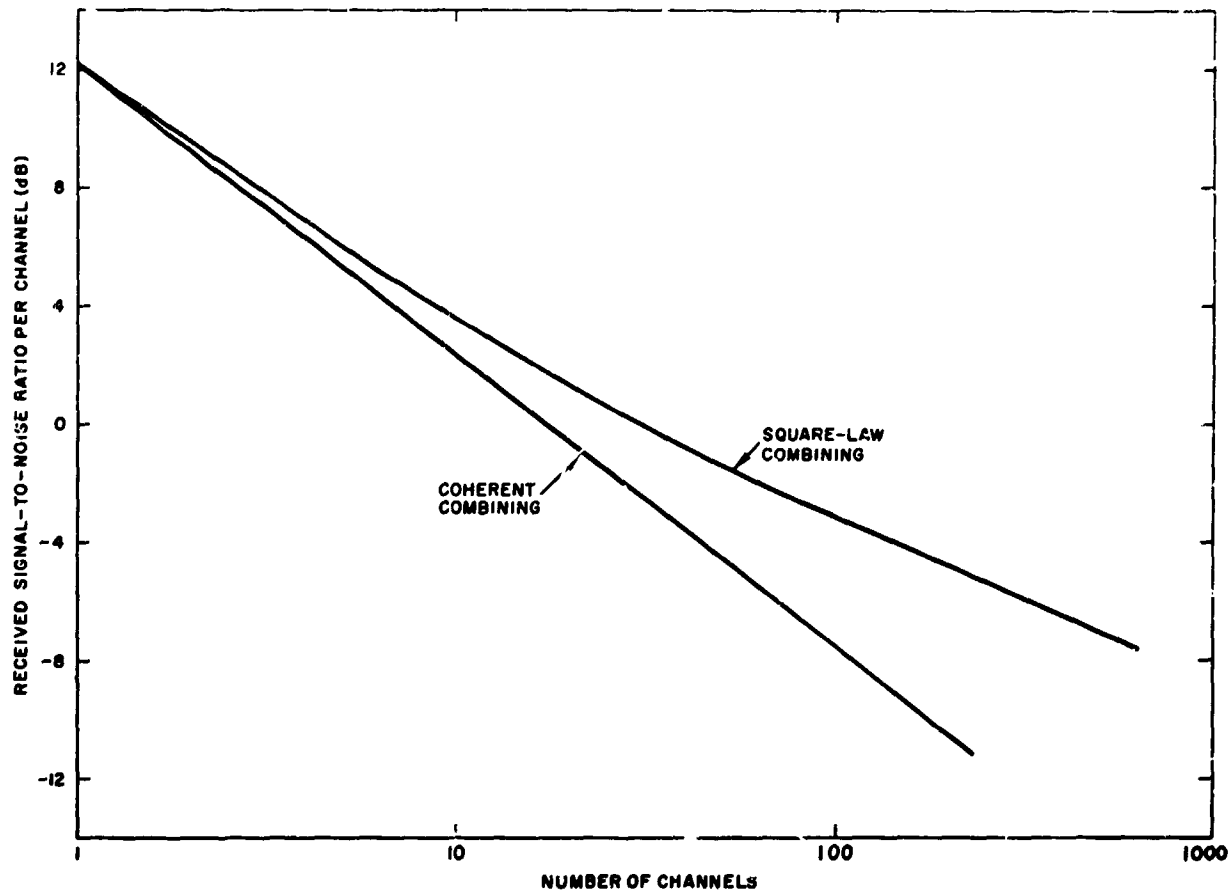


Figure 12-2. Comparison of square law and coherent combining

REFERENCES

1. Private communication with J. N. Lahti, Bell Telephone Laboratories.

2. D. L. Fried, "Optical Heterodyne Detection of an Atmospherically Distorted Signal Wavefront," Proc. IEEE, Vol. 55 (January 1967), pp 57–67.

3. A. E. Siegman, "The Antenna Properties of Optical Heterodyne Receivers," Proc. IEEE, Vol. 54 (October 1966), pp 1350–1356.

4. C. W. Helstrom, Statistical Theory of Signal Detection (New York, Pergamon Press, 1960), Chapter 1.

5. C. C. Cutler, R. Kompfner, and L. C. Tillotson, "A Self-Steering Array Repeater," BSTJ, Vol. 42 (September 1963), pp 2013–2032.